ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746643

ENHANCING AFFECTIVE REPRESENTATIONS OF MUSIC-INDUCED EEG THROUGH MULTIMODAL SUPERVISION AND LATENT DOMAIN ADAPTATION

Kleanthis Avramidis[1,2] *Christos Garoufis*[2] *Athanasia Zlatintsi*[2] *Petros Maragos*[2]

[1] Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA 90089, USA

[2] School of ECE, National Technical University of Athens, 15773 Athens, Greece

ABSTRACT

The study of Music Cognition and neural responses to music has been invaluable in understanding human emotions. Brain signals, though, manifest a highly complex structure that makes processing and retrieving meaningful features challenging, particularly of abstract constructs like affect. Moreover, the performance of learning models is undermined by the limited amount of available neuronal data and their severe inter-subject variability. In this paper we extract efficient, personalized affective representations from EEG signals during music listening. To this end, we employ music signals as a supervisory modality to EEG, aiming to project their semantic correspondence onto a common representation space. We utilize a bi-modal framework by combining an LSTM-based attention model to process EEG and a pre-trained model for music tagging, along with a reverse domain discriminator to align the distributions of the two modalities, further constraining the learning process with emotion tags. The resulting framework can be utilized for emotion recognition both directly, by performing supervised predictions from either modality, and indirectly, by providing relevant music samples to EEG input queries. The experimental findings show the potential of enhancing neuronal data through stimulus information for recognition purposes and yield insights into the distribution and temporal variance of music-induced affective features.

***Index Terms*—** Music Cognition, Emotion Recognition, Electroencephalography, Cross-Modal Learning

1. INTRODUCTION

Music is an abstract, yet densely emotional form of art. It is universally enjoyed, due to its ability to induce powerful emotions irrespective of the underlying mood [1] and has been characterized to greatly affect the function of the human brain [2, 3]. Hence it is widely used to study emotion recognition, both by analyzing the mood produced by several musical features [4, 5] and by studying its effects on human neural and physiological responses [6]. However, the task of extracting emotional information from brain activity poses severe challenges, due to the inherently abstract nature of the induced emotions, the variability in emotional and physiological responses between different individuals and the lack of large-scale databases of emotionally coordinated neural activity. In this paper, we propose a deep multimodal approach [7], using musical stimuli.

The scope of this study lies at the intersection of Music Cognition, Emotion Recognition from neuronal signals and Multimodal Learning. We choose to study brain responses to music by employing a cross-modal system to identify the correspondence between these modalities. We use the electroencephalogram (EEG) to model brain responses for this task and we constrain the learning process with emotion labels. Therefore, we aim to derive important insights regarding the affective role that music can play on humans and the extent to which it can help us build richer neuronal representations of affect. To conduct the experiment, we exploit multimodal optimization and domain adaptation strategies to project EEG and music features onto a common latent space, from which we could assess their similarity. By conditioning the learning process with emotion tags, the constructed space represents affect, enabling thus emotion recognition both directly, by performing supervised predictions, and indirectly, by ranking music tracks to EEG inputs, based on their distance. To the best of our knowledge, this is the first study to propose such a framework, thus it could be utilized as a baseline reference. We also perform an extensive qualitative study across 32 subjects of the DEAP dataset [8] to derive inter-subject affective patterns.

The remainder of this paper is organized as follows: Section 2 reviews the related work in cross-modal learning and research on EEG processing and music cognition. In Section 3 we introduce our problem and present the proposed framework along with the optimization methods we utilized. Section 4 includes information about the data, their pre-processing and implementation details, while in Section 5 we provide the experimental results. In that Section and in Section 6 we provide extensive quantitative and qualitative analysis of the outcomes, while Section 7 concludes our study.

2. RELATED WORK

Music Cognition: Studying the human brain's responses to music stimuli has always been a lively field of research in neuroscience and signal processing [9] aiming to answer fundamental questions regarding our enjoyment of music. The field has gained a lot of attention in recent years, with the upsurge in available neuronal data. Many studies in the field rely on EEG recordings as they provide better temporal resolution than other techniques, such as functional magnetic resonance imaging (fMRI). In addition to the traditional, well-controlled auditory experiments, modern approaches gather physiological data from music listeners as they enjoy or imagine naturalistic music [10], in order for instance to examine correlations in temporal structure [11] or the perceived tempo [12]. One of the core findings on music cognition is the correlation between characteristics of the neural oscillation patterns and rhythmical patterns in music [13]. Additionally, Event-Related Potentials (ERP) have been utilized to extract brain activity patterns that can relate to note onsets or pitch [14, 15]. In parallel to the above, there has also been a shift towards deep learning approaches for information retrieval from music stimuli [16], in which we focus on the present study.

Emotion Recognition: Undoubtedly, the most powerful impact of music on humans concerns the induced emotions. Emotion Recognition is a widely researched field of contemporary Machine Learning and Behavioral Signal Processing [17] and several studies have focused on the musical features [5, 18] that determine affective attributes of music listening in a wide range of emotions.

[2]This research work was performed at NTUA.

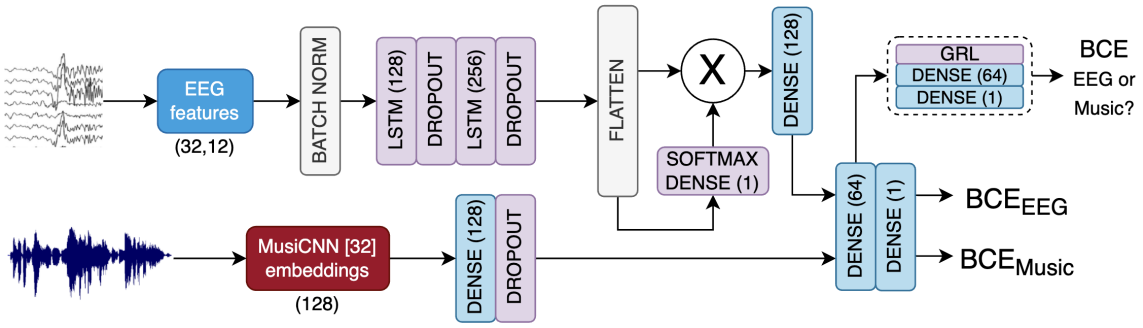

Fig. 1. The proposed bi-stream network. The output embedding layer of each stream is fed to the common 64D dense layer (common space).

Recently, several studies have examined physiological signals to analyze humans' felt emotions [19], with music emerging as an efficient method to elicit them. Due to its temporal resolution, the electroencephalogram is the most widely researched signal of this type and various statistical, spectral or time-frequency features have been proposed for Emotion Recognition [20, 21]. Due to the noisy structure of EEG signals, many studies incorporate entropy [22] and fractal [23] algorithms to extract emotion-related features. Of course, variations of deep neural networks have been proposed and exceeded the performance of traditional feature extraction methods [24, 25], however the limited data availability and inter-subject variability present serious barriers for this kind of modeling.

Cross-Modal Learning: The task of learning a shared embedding space from different datasets or modalities is being studied through various approaches, which are predominantly applied to image and text modalities. A widely used baseline is Canonical Correlation Analysis (CCA). CCA is non-probabilistic and enables the extraction of linear components to optimize the correlation of pairs of vectors. One can find in the literature various non-linear CCA-based frameworks and architectures utilized to learn inter-modal similarities, such as Deep CCA [26]. Besides CCA, other methods that have been used include an HGR-based maximal correlation metric [27] and adversarial training [28], focusing mainly on the optimization function of the respective model, and on adaptive hidden layers [29]. Another study [30] incorporated music to co-train a shared space with images using a contrastive loss. Further, in [31] a state-of-the-art framework exploits label supervision to better manipulate the latent space, a key concept that we also follow in our study.

3. METHODOLOGY

In this study, we extract the semantic relationship between music tracks and corresponding EEG recordings, so that an EEG could be mapped to an efficient affective representation and retrieve emotionally consistent music samples. Let us assume a collection of n instances of EEG-music pairs, denoted as $T = \{(x_i^a, x_i^b)\}_{i=1}^n$ where x_i^a is the input EEG sample of the ith instance and x_i^b the input music stimulus corresponding to that sample. Each instance has been assigned an affective annotation $y_i \in \mathbb{R}^2$ for valence and arousal dimensions. For each instance i we aim to learn an EEG embedding $u_i = f(x_i^a, Y^a) \in \mathbb{R}^d$ and a musical audio embedding $v_i = g(x_i^b, Y^b) \in \mathbb{R}^d$, where d is the dimensionality of the common representation space and Y^a, Y^b the trainable parameters.

3.1. The Proposed Framework

We use a bi-stream Neural Network with one branch for each modality. The EEG branch is a recurrent network, comprised of two LSTM modules and a softmax attention layer. The model takes as input an EEG trial of shape (channels, features) and attempts to capture its inter-channel correlations. Next, the output features of all channels are flattened and passed through an attention module to identify the most important components, that will lead the embedding vector in the common space. We utilize a lightweight network in order to avoid overfitting to the limited range of the available data, however any state-of-the-art model in the task could be applied. For the music branch we use the MusiCNN model [32] to extract high-level embeddings from the available audio stimuli. MusiCNN is a robust network, pre-trained on large audio databases, and produces high-quality music embeddings that compensate for the limited size of our track set and further assist the learning process. The extracted embeddings are then fed into a feed-forward neural network.

To construct the final bi-modal framework, we connect the last layer of each of the previous networks to a dense layer (Fig. 1) constituting the common representation space, from which we output emotion predictions. Inspired by [25], we further apply a Gradient Reversal Layer (GRL) [33], aiming to reduce the distribution shift between EEG and music modalities. In specific, both 64D embeddings are fed to this layer, from where we output a prediction regarding the modality type. From each batch, we randomly permute half of the EEG samples and their respective music samples, forming a new equal-sized batch that we shuffle and input to the GRL module, along with a binary label vector to denote the modality. Subsequently, these embeddings are passed through dense layers to predict the modality of each sample. By reversing the gradients corresponding to these predictions during back-propagation, we help the feature extractor produce modality-invariant features.

3.2. Objective Function

Our goal is to learn a common space where the samples from the same semantic category should be similar, even though they come from different modalities. To learn discriminative features we want to minimize the discrimination loss in both the label space and the representation space, by reducing the cross-modal discrepancy. With regard to the label space, we use a linear classifier to predict the emotion labels of the samples projected in the common space. Outputs of each modality are passed through a sigmoid activation and a binary cross-entropy (BCE) loss ℓ is computed. For the cross-modal task we apply a weighted linear combination of those losses: $\mathcal{J}_1 = \lambda_{11}\ell_a + \lambda_{12}\ell_b$. To reduce the cross-modal discrepancy between EEG and music representations, we also compute the BCE loss of the modality prediction after GRL: $\mathcal{J}_2 = \ell_{dd}$. By combining terms $\mathcal{J}_1, \mathcal{J}_2$ we obtain the proposed objective, in which the hyper-parameters λ_i control the contribution of each separate component and are determined through trial and error: $\mathcal{J} = \lambda_1\mathcal{J}_1 + \lambda_2\mathcal{J}_2$.

4. EXPERIMENTAL SETUP

4.1. The DEAP Dataset

DEAP [8] is a comprehensive dataset that includes EEG signals of music listening, collected from 32 subjects. Each subject watches forty 1-min long music videos while having their EEG recorded. After each video trial, the subject was instructed to rate the emotion that was elicited upon the entire trial in 5 dimensions: valence, arousal, dominance, liking and familiarity to the track. In this paper we solely experiment with the 2D emotion space, determined by valence and arousal, whose ratings range from 1 (weakest) to 9 (strongest). We use the EEG signals in their already preprocessed form: recorded at a sampling rate of 512 Hz and denoised by bandpass filtering, after downsampling to 128 Hz. Eye-related artefacts were removed whereas the 10-20 electrode placement system was followed.

Specifying Music Tracks: The 40 one-minute music stimuli of DEAP are not included in the dataset, so we located the video clips of the corresponding tracks and isolated the minute of interest for each one, according to the metadata provided. The task of deriving the common latent space poses a crucial challenge: the semantic gap between the "subjective" affective responses of participants and the emotion tags of the songs. Ideally, we need musical stimuli that are tagged in accordance with the participants' annotations. DEAP stimuli have been selected for this purpose and have been independently annotated by the experimenters at track level. Nearly all songs received average ratings from the participants that were in accordance with those annotations. We found that only $6/40$ songs had such an inconsistency and discarded them. The resulting track set is used to extract MusiCNN embeddings which we make available[1] as well.

4.2. Input Feature Extraction

EEG and music signals are processed differently in order to produce an embedding form suitable for multimodal training. DEAP signals are first cut to 3-second segments, while having their preparatory phase removed. For feature extraction purposes we consider differential entropy features, reported to achieve superior performance in the task [34]. Differential entropy (DE) h is defined as:

$$h(X) = -\int_X f(x)\log(f(x))dx, \tag{1}$$

where X is an EEG segment and $f(x)$ its distribution. Assuming further that the utilized signals can be modeled as Gaussian distributions, i.e. $f(x) = N(\mu, \sigma)$, then $h(X)$ can be determined by the logarithm energy spectrum of X as follows:

$$\begin{aligned} h(X) = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\frac{(x-\mu)^2}{2\sigma^2} \\ \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\frac{(x-\mu)^2}{2\sigma^2}\right) dx = \frac{1}{2}\log 2\pi e\sigma^2. \end{aligned} \tag{2}$$

Thus, for each EEG segment we use the Short-Time Fourier Transform with an 1-sec non-overlapping Hanning window to compute the variance σ^2 for each of the three windows in the frequency domain and subsequently we compute h in each channel of the four available EEG rhythms: θ (4-7Hz), α (7-13Hz), β (13-30Hz) and γ (31-50Hz). The features for all four bands are concatenated and the resulting feature vector is then used as channel-wise input. On the other hand, music tracks are cut in segments aligned with the EEG throughout their whole (3 sec) duration and fed directly into the pretrained model, from which we extract its "pool5" embeddings.

[1] https://github.com/klean2050/EEG_CrossModal

Dimension	Non-Aggregated	Aggregated
Valence	62.9% – 71.5%	**70.4**% – **78.7**%
Arousal	63.3% – 88.0%	**68.9**% – **91.9**%

Table 1. Emotion Accuracy Scores for (EEG – Music) modalities, reporting mean values over 32 subject-specific models.

Dimension	Precision@10	mAvg. Precision
Valence	**19.4**% – **63.8**%	18.8% – 59.1%
Arousal	18.4% – 65.0%	**19.9**% – **67.8**%

Table 2. (Track – Emotion) Retrieval Scores on EEG input queries, reporting mean aggregated scores over 32 subjects.

4.3. Evaluation Protocol

We evaluate our proposed method using accuracy to assess the supervised predictions for each modality and the Precision@10 (P@10) and mean Average Precision (mAP) metrics for the retrieval of music tracks given EEG queries. Those two metrics have been widely used to assess retrieval tasks in the literature [35, 31] as they evaluate the response's distance-based ranking to each query. In particular, P@10 considers the top 10 ranked tracks whereas mAP evaluates the whole ranking. Results are also presented after trial aggregation: For the accuracy, we simply denote a prediction as correct by majority voting on the segment-wise predictions. For the retrieval metrics, since no such voting can be made, we consider the median of the segment-wise distance scores as the overall query score.

5. EXPERIMENTS

The training procedure considers personalized models, each one trained on data of a specific subject and the respective audio stimuli. To compensate for possible annotation noise, we binarize the emotion labels by setting the threshold to the median score 5, as in [8]. Following the same paradigm, we consider separate experiments for valence and arousal dimensions. We apply 5-fold stratified cross-validation to train each network, where each fold holds 20% of the total trials (7 tracks). Additionally, we apply class weights to alleviate any subject-specific data imbalance. All networks are optimized using Adam at a 10^{-4} learning rate and patience of 15 epochs of non-decreasing validation loss.

5.1. Predicting Emotion Tags

We first evaluate the models' performance on Emotion Recognition for both EEG and Music modalities (Table 1). EEG scores show high variance per subject, on average reaching up to 70.4% on valence and 68.9% on arousal after trial aggregation. The obtained scores are competitive for the specific dataset, despite the simple utilized architecture, something we attribute to the impact of music co-training and the adaptation of the common latent space. This contribution is further quantified in Section 5.3. Additionally, aggregating predictions on a per-track basis provides substantially enhanced results compared to non-aggregated ones, with the EEG accuracy increasing by above 5% in arousal recognition and about 8% in valence, implying that there is strong correlation (e.g., in the form of clusters) between same-track samples, especially in valence. On the other hand, despite the small number of tracks in our music set, the recognition performance is substantially high for the music branch, 78.8% average on valence and 91.9% on arousal, something that indicates the robustness of our transfer learning module.

5.2. Retrieving Tracks from EEG Queries

Table 2 summarizes the retrieval scores from the personalized models, acquired by querying the common representation space of each

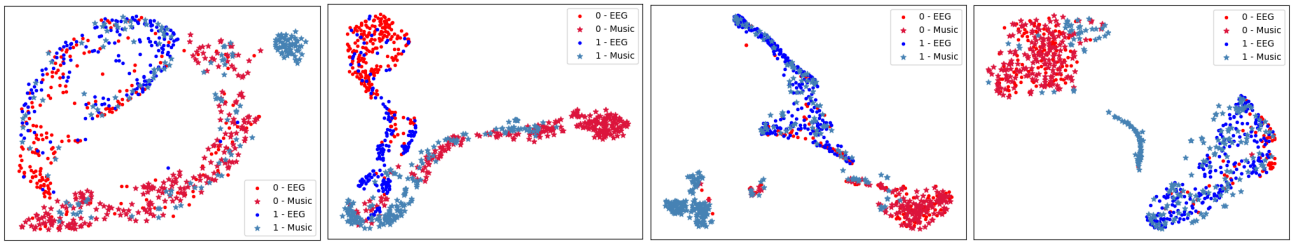

Fig. 2. t-SNE visualisation of the common space for subjects –from left to right– 8, 15 (Valence) and 18, 20 (Arousal). $0 \rightarrow$ Low | $1 \rightarrow$ High

Metric	$\mathcal{J}$	ℓ_a only	$\neg\, \ell_{dd}$
Acc_{EEG}	**70.4**% – **68.9**%	67.8% – 68.0%	67.9% – 63.4%
P@10	**63.8**% – 65.0%	57.3% – 53.1%	63.4% – **66.7**%
mAP	59.1% – 67.8%	51.9% – 55.8%	**59.8**% – **68.1**%

Table 3. Ablation on the Objective Function for (Valence – Arousal). Here we solely consider mean aggregated scores over 32 subjects.

network with a test EEG sample and then evaluating the ranking of music samples based on their distance to the query. Retrieval metrics provide robust results in both cases, indicating that the EEG samples are well-situated in the common space and the majority of them are capable of retrieving tracks that are emotionally consistent. Specifically, in the case of induced valence, a P@10 value of 63.8% is achieved. We note that this percentage is higher than the reported mAP (59.1%), strongly implying that the learned valence space is fragmented into local subspaces of high similarity. Arousal on the other hand seems to be more consistently represented, as both mAP and P@10 median retrieval scores indicate that the majority of tested tracks can derive consistent music rankings, in contrast to valence where the emotional response similarity seems concentrated to the top-ranked elements. As a result, the correct retrieval percentage conditioned on arousal approaches 68% on average across subjects. We also note some preliminary results in approaching retrieval of the exact stimulus of an EEG sample. The derived scores, around 20%, are clearly above random selection, however we believe that further experimentation is required on the temporal resolution of the input samples, yielding an interesting direction of future study.

5.3. Ablation Study

In our study we incorporated a complex objective function, combining 3 BCE terms to minimize the discrimination loss in both the label space and the common latent space. To further investigate the impact of our proposals on the models' performance, we trained separate sessions, first by considering sole EEG samples without music supervision, and second by avoiding the domain discrimination module. From the results in Table 3 we deduce that our full objective $\mathcal{J}$ leads to higher overall performance, indicating that all utilized terms contribute to richer EEG affective representations. Specifically, we can see that the absence of multimodal training sharply impacts the validity of the common space and reduces the classification performance, 2.6% in valence and 0.9% in arousal. On the other side, the absence of domain adaptation causes slighter modifications to the correlation of samples and stimuli, as measured by precision metrics. Through this module we manage though to better distribute samples in the common space, break modality-specific clusters and reduce the overall sample distances (Section 6.1), which is reflected in the improved classification performance in both experiments.

6. QUALITATIVE ANALYSIS

Studying the Common Space: We visually inspect the produced latent space using t-SNE to reduce its 64D dimension to 2D. We select one of the 5 trained models for 2 subjects in Valence and Arousal to display their results in Fig. 2 (similar trends are observed for most subjects). It is evident that latent domain adaptation has alleviated the cross-modal discrepancy and the modalities homogenize their embeddings to a certain degree. Cohesive sub-clusters are though visible, especially in the case of valence. This provides an explanation towards the discrepancy we observed between P@10 and mAP metrics, since the top-ranked track retrievals originate from the corresponding local subspace, but there is no coarse bisection between high- and low- valence samples, in contrast to the case of arousal.

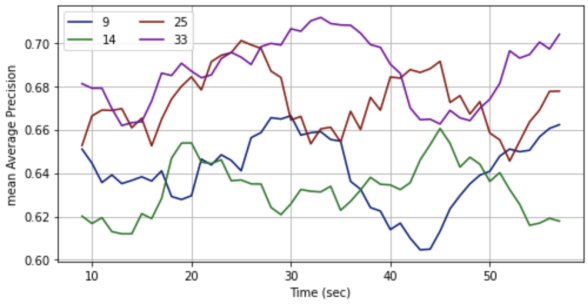

Fig. 3. Arousal mAP scores over the 58 time samples for the numbered tracks, averaged across all subjects.

Temporal Variation of Recognition: Since each track is segmented into 58 overlapping samples of 3 sec, it is expected that the emotion is not elicited at the same pace throughout its duration. Hence, the temporal variation of our scores could indicate important moments in the track. In Fig. 3, we present the temporal evolution of the mAP scores for selected music tracks, averaged across all subjects. While the raw plots are noisy, each song individually exhibits a pattern of variation, which we depict by applying a 7-sample moving average filter. Scores typically reveal an oscillating pattern on the time axis and emotions are highly induced at certain peaks of the graph. These patterns reveal a characteristic picture of emotional induction in songs and could be subject of further experimentation.

7. CONCLUSION

In this paper we presented a novel approach to analyze emotion induction from EEG recordings of music listening. We proposed a cross-modal framework to learn rich affective representations for EEG data through music supervision and adaptation of a common latent space, from which one could retrieve consistent music rankings from EEG queries. Our approach indicates that distilling information from processed musical stimuli to the respective EEG signals can improve performance and provide insights on personalized emotion analysis. To the best of our knowledge, this is the first study to propose a complete framework for the specific task and dataset, thus our results can be viewed as a concrete baseline. This framework can be used to model the EEG-Music relationship by using different condition mechanisms, e.g., musical beat. Another interesting direction would be to explore improvements in exact stimulus retrieval.

8. REFERENCES

[1] K. Jakubowski and A. Ghosh, "Music-evoked autobiographical memories in everyday life," *Psychology of music*, vol. 49, no. 3, pp. 649–666, 2021.

[2] O. W. Sacks, *Musicophilia : Tales of Music and the Brain*, Alfred A. Knopf, New York, 1st edition, 2007.

[3] D. Bashwiner, "Brain and Music. By Stefan Koelsch," *Music Theory Spectrum*, vol. 39, no. 2, pp. 269–274, 2017.

[4] R. Panda, R. Malheiro, and R. P. Paiva, "Novel Audio Features for Music Emotion Recognition," *IEEE Trans. Affective Computing*, 2020.

[5] Y. Song, S. Dixon, and M. Pearce, "Evaluation of Musical Features for Emotion Classification," in *Proc. ISMIR 2012*, Porto, Portugal, 2012.

[6] T. Greer, B. Ma, M. Sachs, A. Habibi, and S. Narayanan, "A Multimodal View into Music's Effect on Human Neural, Physiological, and Emotional Experience," in *Proc. ACM Int'l Multimedia Conf. 2019*, Nice, France, 2019.

[7] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. ICML 2011*, Bellevue, WA, USA, 2011.

[8] S. Koelstra et al., "DEAP: A Database for Emotion Analysis Using Physiological Signals," *IEEE Trans. Affective Computing*, vol. 3, 2011.

[9] T. Schäfer, P. Sedlmeier, C. Städtler, and D. Huron, "The psychological functions of music listening," *Frontiers in Psychology*, vol. 4, pp. 511, 2013.

[10] S. Losorelli, D. T. Nguyen, J. Dmochowski, and B. Kaneshiro, "NMED-T: A Tempo-Focused Dataset of Cortical and Behavioral Responses to Naturalistic Music," in *Proc. ISMIR 2017*, Suzhou, China, 2017.

[11] A. Vinay, A. Lerch, and G. Leslie, "Mind the Beat: Detecting Audio Onsets from EEG Recordings of Music Listening," *ArXiv*, vol. abs/2102.06393, 2021.

[12] S. Stober, T. Prätzlich, and M. Müller, "Brain Beats: Tempo Extraction from EEG Data," in *Proc. ISMIR 2016*, New York, NY, USA, 2016.

[13] S. Nozaradan, I. Peretz, M. Missal, and A. Mouraux, "Tagging the Neuronal Entrainment to Beat and Meter," *Journal of Neuroscience*, pp. 10234–10240, 2011.

[14] R. S. Schaefer, P. Desain, and P. Suppes, "Structural Decomposition of EEG Signatures of Melodic Processing," *Biological Psychology*, pp. 253–259, 2009.

[15] H. Poikonen, et al., "Event-related Brain Responses while Listening to Entire Pieces of Music," *Neuroscience*, vol. 312, pp. 58–73, 2016.

[16] E. J. Humphrey, J. P. Bello, and Y. LeCun, "Moving Beyond Feature Design: Deep Architectures and Automatic Feature Learning in Music Informatics.," in *Proc. ISMIR 2012*, Porto, Portugal, 2012.

[17] M. Sarprasatham, "Emotion Recognition: A Survey," *Int'l Journal of Advanced Research in Computer Science*, vol. 3, pp. 14–19, 01 2015.

[18] R. Panda, B. Rocha, and R. P. Paiva, "Music Emotion Recognition with Standard and Melodic Audio Features," *Applied Artificial Intelligence*, pp. 313–334, 2015.

[19] S. Chen, Z. Gao, and S. Wang, "Emotion recognition from peripheral physiological signals enhanced by EEG," in *Proc. ICASSP 2016*, New Orleans, LA, USA, 2016.

[20] X. Wang, D. Nie, and B. Lu, "EEG-Based Emotion Recognition Using Frequency Domain Features and Support Vector Machines," in *Proc. ICONIP 2011*, Shanghai, China, 2011.

[21] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion Recognition from Brain Signals Using Hybrid Adaptive Filtering and Higher Order Crossings Analysis," *IEEE Trans. Affective Computing*, vol. 1, no. 2, pp. 81–97, 2010.

[22] R. Duan, J. Zhu, and B. Lu, "Differential Entropy Feature for EEG-based Emotion Classification," in *Proc. Int'l IEEE/EMBS Conf. on Neural Engineering (NER)*, San Diego, CA, USA, 2013.

[23] K. Avramidis, A. Zlatintsi, C. Garoufis, and P. Maragos, "Multiscale Fractal Analysis on EEG Signals for Music-Induced Emotion Recognition," in *Proc. EUSIPCO 2021*, Amsterdam, the Netherlands, 2021.

[24] Y. Wang, Z. Huang, B. McCane, and P. Neo, "EmotioNet: A 3-D Convolutional Neural Network for EEG-based Emotion Recognition," in *Proc. IJCNN 2018*, Rio de Janeiro, Brazil, 2018.

[25] X. Du et al., "An Efficient LSTM Network for Emotion Recognition from Multichannel EEG Signals," *IEEE Trans. Affective Computing*, pp. 1–1, 2020.

[26] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep Canonical Correlation Analysis," in *Proc. of Machine Learning Research*, Atlanta, GA, USA, 2013, pp. 1247–1255.

[27] M. Li, Y. Li, S.-L. Huang, and L. Zhang, "Semantically Supervised Maximal Correlation For Cross-Modal Retrieval," in *Proc. ICIP 2020*, Abu Dhabi, UAE, 2020.

[28] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial Cross-Modal Retrieval," in *Proc. ACM Int'l Multimedia Conf. 2017*, Mountain View, CA, USA, 2017.

[29] T.-K. Hu, Y.-Y. Lin, and P.-C. Hsiu, "Learning Adaptive Hidden Layers for Mobile Gesture Recognition," in *Proc. AAAI 2018*, New Orleans, LA, USA, 2018.

[30] B. Li and A. Kumar, "Query by Video: Cross-modal Music Retrieval," in *Proc. ISMIR 2019*, Delft, the Netherlands, 2019.

[31] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep Supervised Cross-Modal Retrieval," in *Proc. CVPR 2019*, Salt Lake City, UT, USA, 2019.

[32] J. Pons, O. Nieto, M. Prockup, E. M. Schmidt, A. F. Ehmann, and X. Serra, "End-to-End Learning for Music Audio Tagging at Scale," in *Proc. ISMIR 2018*, Paris, France, 2018.

[33] Y. Ganin and V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," in *Proc. ICML 2015*, Lille, France, 2015.

[34] W.-L. Zheng and B.-L. Lu, "Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks," *IEEE Trans. on Autonomous Mental Development*, vol. 7, pp. 162–175, 2015.

[35] M. Won, S. Oramas, O. Nieto, F. Gouyon, and X. Serra, "Multimodal Metric Learning for Tag-based Music Retrieval," *arXiv preprint:2010.16030*, 2020.