2019 IEEE International Conference on Cognitive Computing (ICCC)

Fusing Acoustic and Electroencephalographic Modalities for User-Independent Emotion Prediction

Stavros Ntalampiras
Department of Computer Science
University of Milan
Milan, Italy
name.surname@unimi.it

Federico Avanzini
Department of Computer Science
University of Milan
Milan, Italy
name.surname@unimi.it

Luca Andrea Ludovico
Department of Computer Science
University of Milan
Milan, Italy
name.surname@unimi.it

***Abstract*—Search and retrieval of multimedia content based on the evoked emotion comprises an interesting scientific field with numerous applications. This paper proposes a method that fuses two heterogeneous modalities, i.e. music and electroencephalographic signals, both for predicting emotional dimensions in the valence-arousal plane and for addressing four binary classification tasks, namely i.e. high/low arousal, positive/negative valence, high/low dominance, high/low liking. The proposed solution exploits Mel-scaled and EEG spectrograms feeding a k-medoids clustering scheme based on canonical correlation analysis. A thorough experimental campaign carried out on a publicly available dataset confirms the efficacy of such an approach. Despite its low computational cost, it was able to surpass state of the art results, and most importantly, in a user-independent manner.**

***Keywords*-music emotion prediction; EEG emotion prediction; music EEG fusion; canonical correlation analysis; k-medoids clustering algorithm.**

I. Introduction

Emotions possess a role of paramount importance in communication between humans and our interpersonal relationships as they have an effect on our intelligence and influence our thoughts [1], [2]. In the early years of widespread usage of computer systems, machines were not required nor expected to have any emotional understanding. However, due to the advances on a series of scientific fields including computer science, neuroscience, and psychology, machines are quickly gaining such emotion sensing skills. In this context, one interesting application lies within personalized recommendation systems, where a computer is able to learn on-the-fly the emotions expressed by its user, and readjust its functionalities to better fit his/her needs. This can be seen as the primary step towards realizing a profound element of intelligence in human-computer interaction [3], [4].

Several paths exist towards realizing the goals of affective computing, i.e. minimizing the distance between the end-user and computer systems via emotionally perceptive human computer interaction [5]. This paper employs two such paths, i.e. electroencephalographic (EEG) signals and music content. The former has recently emerged [6] in the context of so-called affective brain-computer interfaces (aBCIs). BCIs are commonly used for communication and rehabilitation purposes [7], while aBCIs incorporate user's emotional state in their operation loop so as to improve their functionality. Interesting applications include neurogaming [8], neuromarketing [9], the so-called attention monitors [10], and so on. The most prominent modality in aBCIs is electroencephalography since *a)* it is non-invasive, *b)* it is portable, *c)* it offers high temporal resolution, and *d)* it has acceptable costs [11].

This work concentrates on predicting the emotions elicited on users while listening to music towards automated affective music tagging. The latter source of information exploited within affective computing is the acoustic one, where essentially an audio signal is associated with the emotion evoked on its listener. The overall aim is to model that relationship, thus being able to predict the emotional responses of music signals [12]–[14]. Motivated by the diverse characteristics of EEG and music signals, this work explores their simultaneous usage for emotion prediction.

II. Related work

The literature includes several monomodal approaches, i.e. using either EEG or music signals, addressing the specific problem. Emotion-prediction solutions using solely EEG signals exist in the literature. Özerdem and Polat [15] employ wavelet transform along with multilayer perceptron neural network and k-nearest neighborhood algorithm to classify EEG signals. They concentrate on classifying between positive (valence$\geq$ 5) and negative (valence$<$ 5) emotional states. Clerico *et al.* [16] combine conventional features with EEG amplitude modulations. To address four different binary classification problems, (detecting low/high valence, low/high arousal, low/high dominance and low/high liking), they used support vector machines (SVM), relevance vector machines and random forest classifiers. With regard to the acoustic modality, a recent work [17] provides an interesting review of music emotion recognition based solely on the available audio information. Following the same audio-only line of thought, Aljanaki *et al.* [18] present a benchmark for emotional analysis of music.

978-1-7281-2711-8/19/$31.00 ©2019 IEEE
DOI 10.1109/ICCC.2019.00018

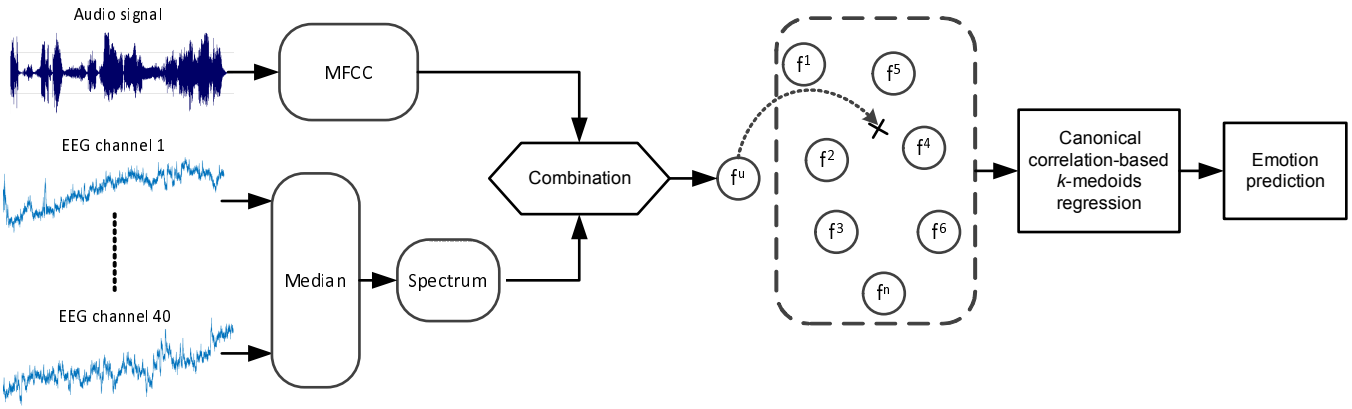

Figure 1. The main idea of this work: after multimodal feature extraction and fusion, a canonical correlation-based k-medoids regression scheme achieves emotion prediction.

Moving to information fusion approaches, a method based on the Dempster-Shafer theory has been recently proposed [19]. The authors use as inputs the audio and video modalities, as well as viewers comments associated with YouTube videos without taking EEG recordings into account. A related work [20] employs the video cues in a hybrid multilevel fusion. There, a binary classification problem is formed, i.e high/low arousal and positive/negative valence. Lin *et al.* [21] present a systematic approach fusing electroencephalography signals and musical contents for classifying valance and arousal. The authors used the Oscar movie soundtrack dataset [22] for extracting EEG and music feature sets, which were then juxtaposed to feed a SVM classifier with a radial basis function. Ultimately, the addressed task was a binary emotion classification problem: positive vs. negative valence and arousal.

This work exploits both the acoustic and EEG modalities to address emotion prediction/classification. To this end, we employ the Database for Emotion Analysis using Physiological Signals (DEAP) dataset [23]. Even though video cues could also be useful, it is possible for a song to be accompanied by various completely different video clips, thus resulting in diverse video-derived feature distributions. As an example, several video cues existing on YouTube have changed since the creation of the DEAP dataset making the video modality useless when it comes to compare with state-of-the-art methods employing the video signal. However, music content and EEG signals are consistent, thus this work relies on those modalities.

Figure 1 summarizes the approach followed in this work. After extracting two different types of spectrum, we propose to use a k-medoids clustering scheme based on canonical correlation analysis able to address both emotion prediction and classification. Thorough experiments demonstrate the efficiency of such a solution which is able to provide encouraging performance in a user-independent setting.

III. The Proposed Solution for Emotion Prediction

This section analyzes the proposed solution for emotion prediction via the simultaneous consideration of music and EEG signals. In the beginning, we explain the feature extraction process associated with signals of different modalities. Moving on, we detail the feature fusion process and finally the k-medoids regression that realizes emotion prediction.

A. Feature extraction

The features extracted out of the audio and EEG signals are described next.

1) Mel-Frequency Cepstral Coefficients: For the derivation of the first feature set, 23 Mel filter bank log-energies are utilized. The extraction method is the following: firstly, the Discrete Fourier transform (DFT) is computed for every frame, and its outcome is filtered using a triangular Mel scale filterbank. Subsequently, we compute the logarithm to adequately space the data. At this point, the discrete cosine transform is employed to reduce the dimensionality and decorrelate the obtained vectors. With this procedure, the 23 log-Mel filterbank coefficients are reduced to 13. First and second derivatives are juxtaposed to the main feature vector.

Moreover, each sound is cut into frames of 30 ms with 10 ms overlap for enabling robustness to possible misalignments. The sampled data are Hamming-windowed to smooth any discontinuities while the DFT size is 512.

2) EEG spectrogram: Towards computing the electroencephalogram spectrogram, we first calculate the median value of the 40 EEG channels sampled at 512 Hz [23]. Subsequently, as shown in Fig. 1, we compute its spectrum after cutting the signal into frames of 1024 samples using the Hamming window. Finally, the EEG spectrogram is combined with the MFCCs, thus achieving information fusion at the feature level.

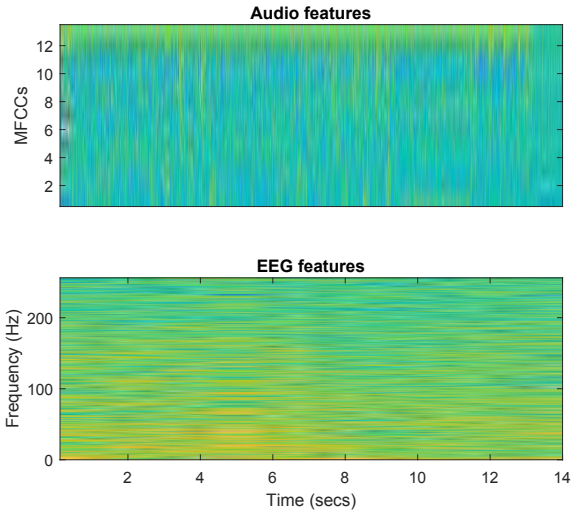

Figure 2. Features extracted out of the song entitled "Jungle Drum", by Emiliana Torrini, along with the corresponding EEG features extracted out of the responses of the first subject.

A representative example of MFCC features extracted from a song, along with the corresponding EEG features extracted from the responses of the first subject, is demonstrated in Fig. 2.

B. k-medoids regression based on canonical correlation

The goal of the proposed regression algorithm is to identify feature vectors characterized by closely-spaced emotional annotations. To achieve this goal, we rely on the *k-mediods clustering algorithm* [24], which belongs to the k-means family. The main characteristic of this algorithm is the ability to consider the more general concept of *pairwise dissimilarity* as distance metric, rather than the Euclidean distance (as done in the traditional k-means algorithm). This may guarantee a more robust clustering phase since the effects of outliers are significantly reduced [24]. Its usage is motivated by its successful application on the sound emotion prediction domain described in [13].

In our specific case, the elements of the k-medoids algorithm are the feature matrices obtained by juxtaposing the MFCCs and the EEG spectrum. As pairwise-dissimilarity measure of the algorithm, we propose the distance metric $\mathcal{D}_{i,j}$ measuring the canonical correlation between vector f^i and f^j, with f^i and f^j being feature matrices coming from different samples in the dataset. Such a metric is able to assess the correlations exhibited by multi-dimensional datastreams, since it provides a way to search and discover linear combinations of the feature vector coefficients exhibiting maximum correlations with each other [25].

In general, when using canonical correlation analysis (CCA), two or more different representations of the same entity (emotion) are available, while the purpose of CCA is to compute a projection for each representation in such a way that they are maximally correlated in the dimensionality-reduced space. Let the feature vectors be f^i and f^j respectively. CCA provides two projection vectors $w_i \in \mathcal{R}^d$ and $w_j \in \mathcal{R}^d$, where d is the dimensionality of the data such that the correlation coefficient ρ is maximized:

$$\rho_{i,j} = \frac{w_i^T f^i f^{j,T} w_j}{\sqrt{(w_i^T f^i f^{i,T} w_i)(w_j^T f^j f^{j,T} w_j)}}. \quad (1)$$

Interestingly, $\rho_{i,j}$ does not depend on the scaling of w_i nor w_j, thus the CCA problem can be formalized as follows:

$$\arg\max_{w_i, w_j} w_i^T f^i f^{j,T} w_j, \quad (2)$$

subject to

$$w_i^T f^i f^{i,T} w_i = w_j^T f^j f^{j,T} w_j = 1. \quad (3)$$

Assuming that $f^j f^{j,T}$ is not singular, it can be shown [26] that the following optimization problem may provide a solution for determining w_i:

$$\operatorname*{argmax}_{w_i} w_i^T f^i f^{j,T} (f^j f^{j,T})^{-1} w_i, \quad (4)$$

subject to

$$w_i^T f^i f^{i,T} w_i = 1. \quad (5)$$

The vector w_j can be computed in a similar way. This provides $\rho_{i,j}$ and $\mathcal{D}_{i,j} = 1/\rho_{i,j}$. Obviously, the lower the values of $\mathcal{D}_{i,j}$, the closer the respective feature vectors in the CCA distance space. We emphasize that $\mathcal{D}_{i,j}$ is symmetric, i.e. $\mathcal{D}_{i,j} = \mathcal{D}_{j,i}$. This process identifies the k closest feature vectors with respect to the unknown one f^u.

Suppose that the emotional content is annotated with arousal, valence, dominance, and liking measurements $a_1, \ldots, a_k,\ v_1, \ldots, v_k,\ d_1, \ldots, d_k,\ l_1, \ldots, l_k,\ a_u,\ v_u,\ d_u$, and l_u respectively. The unknown characteristics of f^u are predicted as their averaged values computed as follows:

$$a_u = \frac{1}{k}\sum_{k=1}^{n} a_k, \quad v_u = \frac{1}{k}\sum_{k=1}^{n} v_k,$$
$$d_u = \frac{1}{k}\sum_{k=1}^{n} d_k, \quad l_u = \frac{1}{k}\sum_{k=1}^{n} l_k.$$

In the proposed emotion prediction system, the implementation of the k-medoids algorithm is based on the Partitioning Around Medoids [27]. The following section explains the experimental set-up and analyses the obtained results.

IV. Experiments

This section details the emotionally annotated music and EEG dataset, and subsequently presents and analyses the prediction and classification results obtained by the proposed solution.

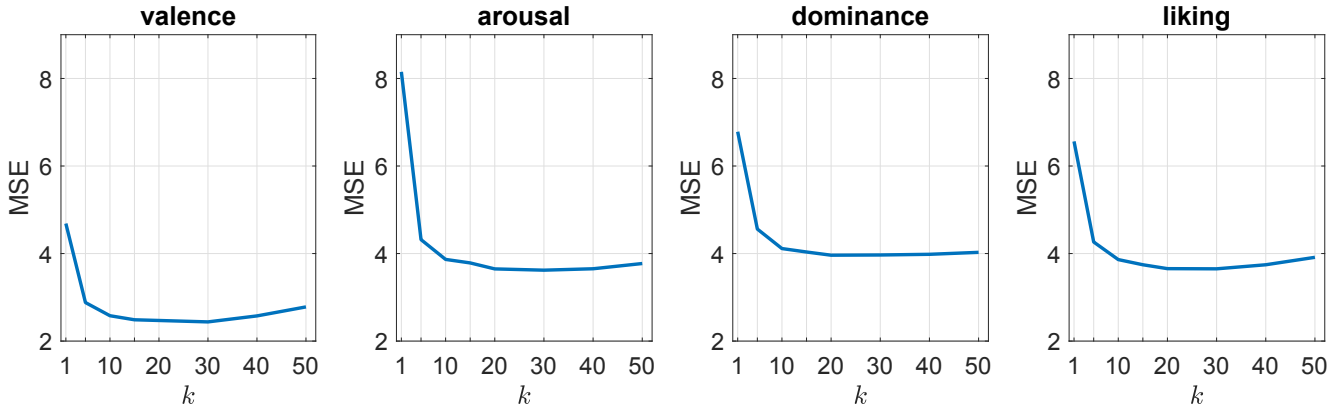

Figure 3. The MSE vs. k obtained by the proposed solution with respect to all emotional dimensions, i.e. valence, arousal, dominance, and liking.

A. Dataset

Both the music and EEG signals employed in this study were taken from the publicly available DEAP dataset.[1] In brief, the dataset is composed of 40 music video clips along with the emotion perceived by 32 subjects as they recorded it via a web-based interface.

Each subject was asked to use a discrete nine-point scale to rate the following quantities: *a) valence* ranging from unhappy/sad to happy/joyful, *b) arousal* ranging from calm/bored to stimulated/excited *c) dominance* ranging from submissive/without control to dominant/in control, empowered, and *d) liking* indicating how much they liked the video clip.

The 40 music signals are divided so that they cover the entire arousal-valence space in a uniform way, i.e. ten songs per quadrant. Moreover, the rating of the selected songs had strong agreement between participants as observed by the respective variation [23].

The selected data were annotated by 32 participants (50% male), aged between 19 and 37 (mean age 26.9) while having their physiological responses recorded. Each subject signed an informed consent prior to the start of the experiment, which conformed with the Ethics of the World Medical Association (Declaration of Helsinki).

Here, we process the music and EEG signals as recorded by the participants while listening to each song. The proposed method can address both the problem of predicting the actual *valence*, *arousal*, *dominance*, and *liking* values as well as four binary classification problems, i.e. high/low arousal, positive/negative valence, high/low dominance, high/low liking, as it is usually done in the related literature [20]. Towards obtaining the ground truth for the classification case, we thresholded the ratings in the middle [20].

[1] http://www.eecs.qmul.ac.uk/mmv/datasets/deap/

B. Results and analysis

Here, we present the results obtained for the two tasks, i.e. the regression one aiming at predicting the exact rating value, and the four binary classification tasks. It should be noted that, towards inferring the class prediction, we followed the thresholding process used for obtaining the ground truth. Moreover, we employed a leave one user out (LOUO) validation scheme, which is a key element of this experimental campaign as one cannot expect to have knowledge regarding ratings of a new subject. As such, a LOUO testing process is deemed more appropriate.

Figure 3 depicts the mean square error (MSE) (its standard formulation as in [28] was employed) vs. k obtained by the proposed solution with respect to all emotional dimensions. We observe that, for all ratings, a value k around 20 neighbors is able to provide the most accurate prediction. All MSEs are between 2 and 4, which demonstrates the efficacy of the proposed k-medoids scheme based on CCA. The best predictions, i.e. the ones with the lowest MSEs, concern the *valence* dimension, while the worst the *dominance* one. Overall, we see that such a correlation-based prediction scheme is able to accurately infer all emotional dimensions in a user-independent manner. Unfortunately, we could not find a related paper aiming to predict the ratings' values to compare with.

Figure 4 demonstrates the classification rates vs. k of the proposed solution with respect to four binary tasks, i.e. high/low arousal, positive/negative valence, high/low dominance, high/low liking. We can see that values of k between 10 and 30 offer the best classification rates in every case. The obtained rates are 83%, 82.4%, 84.7%, and 83% for the valence, arousal, dominance, and liking classification tasks, respectively. Given the low complexity of the proposed method, as it is based on the computation of Mel-scaled and EEG spectrograms, as well as feature vector correlations, we find the achieved rates quite encouraging in validating the assumption that strongly correlated feature vectors share

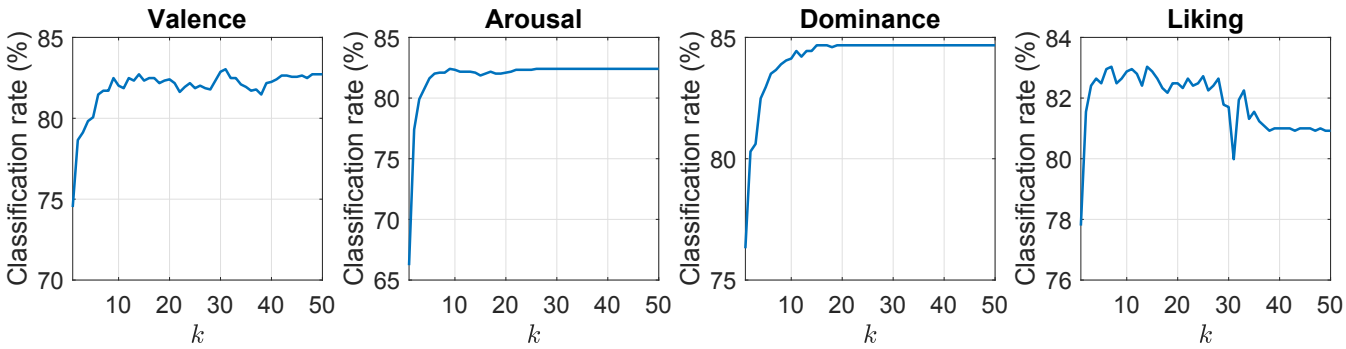

Figure 4. The classification rates vs. k obtained by the proposed solution with respect to all emotional dimensions, i.e. valence, arousal, dominance, and liking.

similar emotional dimensions in the valence-arousal plane. Interestingly, this is the first paper addressing the *liking* classification task.

V. Conclusions

To the best of our knowledge, the highest rates reported in the literature [20] are 90%, 74%, and 72% for the valence, arousal, and dominance classification tasks, respectively, without any consideration of the liking dimension. The results of the present paper are superior only for the arousal and dominance classification tasks. However, these are only indirectly comparable, since they were not calculated under a user-independent setting, while the contrasted solution makes use of the video modality, thus requiring increased computational resources. Conclusively, we infer that the proposed fusion of music and EEG signals is able to offer accurate prediction of emotional ratings in the valence-arousal plane, as well as the involved binary classification tasks.

This work demonstrated that correlations existing in the frequency domain of EEG and audio signals are able to provide satisfying performance for emotion prediction and classification. More importantly, the proposed solution is able to operate in a user-independent setting surpassing state-of-the-art results.

Our future work includes incorporating the emotion prediction mechanism in personalized recommendation systems, and consider its output for emotion-aware suggestions. There, we intend to fully assess its performance, detect its limitations and get clear insights on how to improve its functionality.

References

[1] G. Loewenstein and J. Lerner, *The role of affect in decision making*. Oxford: Oxford University Press, 2003, pp. 619–642.

[2] B. D. Martino, "Frames, biases, and rational decision-making in the human brain," *Science*, vol. 313, no. 5787, pp. 684–687, aug 2006. [Online]. Available: https://doi.org/10.1126/science.1128356

[3] J. Preece, Y. Rogers, H. Sharp, D. Benyon, S. Holland, and T. Carey, *Human Computer Interaction*. Wokingham, England: Addison-Wesley, 1994.

[4] S. Ntalampiras and I. Potamitis, "On predicting the unpleasantness level of a sound event," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 1782–1785. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2014/i14_1782.html

[5] S. Ntalampiras and I. Potamitis, "A statistical inference framework for understanding music-related brain activity," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–1, 2019.

[6] C. Mãhl, C. Jeunet, and F. Lotte, "EEG-based workload estimation across affective contexts," *Frontiers in Neuroscience*, vol. 8, jun 2014. [Online]. Available: https://doi.org/10.3389/fnins.2014.00114

[7] T. W. Kjaer and H. B. Sørensen, "A brain-computer interface to support functional recovery," in *Frontiers of Neurology and Neuroscience*. S. KARGER AG, 2013, pp. 95–100. [Online]. Available: https://doi.org/10.1159/000346430

[8] D. P. O. Bos, B. Reuderink, B. van de Laar, H. Gurkok, C. Muhl, M. Poel, D. Heylen, and A. Nijholt, "Human-computer interaction for BCI games: Usability and user experience," in *2010 International Conference on Cyberworlds*, Oct 2010, pp. 277–281.

[9] N. Lee, A. J. Broderick, and L. Chamberlain, "What is 'neuromarketing'? A discussion and agenda for future research," *International Journal of Psychophysiology*, vol. 63, no. 2, pp. 199–204, feb 2007. [Online]. Available: https://doi.org/10.1016/j.ijpsycho.2006.03.007

[10] M. M. Jackson and R. Mappus, "Applications for brain-computer interfaces," in *Brain-Computer Interfaces*. Springer London, 2010, pp. 89–103. [Online]. Available: https://doi.org/10.1007/978-1-84996-272-8_6

[11] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 327–339, July 2014.

[12] X. Li, H. Xianyu, J. Tian, W. Chen, F. Meng, M. Xu, and L. Cai, "A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 544–548.

[13] S. Ntalampiras, "A transfer learning framework for predicting the emotional content of generalized sound events," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 1694–1701, mar 2017. [Online]. Available: https://doi.org/10.1121/1.4977749

[14] H. Xianyu, X. Li, W. Chen, F. Meng, J. Tian, M. Xu, and L. Cai, "SVR based double-scale regression for dynamic emotion prediction in music," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 549–553.

[15] M. S. Özerdem and H. Polat, "Emotion recognition based on EEG features in movie clips with channel selection," *Brain Informatics*, vol. 4, no. 4, pp. 241–252, jul 2017. [Online]. Available: https://doi.org/10.1007/s40708-017-0069-3

[16] A. Clerico, A. Tiwari, R. Gupta, S. Jayaraman, and T. H. Falk, "Electroencephalography amplitude modulation analysis for automated affective tagging of music video clips," *Frontiers in Computational Neuroscience*, vol. 11, jan 2018. [Online]. Available: https://doi.org/10.3389/fncom.2017.00115

[17] X. Yang, Y. Dong, and J. Li, "Review of data features-based music emotion recognition methods," *Multimedia Systems*, aug 2017. [Online]. Available: https://doi.org/10.1007/s00530-017-0559-4

[18] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PLOS ONE*, vol. 12, no. 3, p. e0173392, mar 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0173392

[19] S. Nemati and A. R. Naghsh-Nilchi, "An evidential data fusion method for affective music video retrieval," *Intelligent Data Analysis*, vol. 21, no. 2, p. 427441, Mar 2017. [Online]. Available: http://doi.org/10.3233/IDA-160029

[20] A. Yazdani, E. Skodras, N. Fakotakis, and T. Ebrahimi, "Multimedia content analysis for emotional characterization of music video clips," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, apr 2013. [Online]. Available: https://doi.org/10.1186/1687-5281-2013-26

[21] Y.-P. Lin, Y.-H. Yang, and T.-P. Jung, "Fusion of electroencephalographic dynamics and musical contents for estimating emotional responses in music listening," *Frontiers in Neuroscience*, vol. 8, may 2014. [Online]. Available: https://doi.org/10.3389/fnins.2014.00094

[22] Y. P. Lin, C. H. Wang, T. L. Wu, S. K. Jeng, and J. H. Chen, "EEG-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 489–492.

[23] 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, Jan 2012.

[24] L. Kaufman and P. Rousseeuw, "Clustering by means of medoids," in *Statistical Data Analysis Based on the L1-Norm and Related Methods*, Y. Dodge, Ed. North-Holland, 1987, pp. 405–416.

[25] *Canonical Correlation Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 321–330. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-72244-1_14

[26] L. Sun, S. Ji, and J. Ye, "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 194–200, Jan 2011.

[27] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, Third Edition*. Orlando, FL, USA: Academic Press, Inc., 2006.

[28] C. Alippi, S. Ntalampiras, and M. Roveri, "Model ensemble for an effective on-line reconstruction of missing data in sensor networks," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, Aug 2013, pp. 1–6.