Improving Induced Valence Recognition by Integrating Acoustic Sound Semantics in Movies

Shreya G. Upadhyay, Bo-Hao Su, Chi-Chun Lee
Department of Electrical Engineering, National Tsing Hua University, Taiwan
shreya@gapp.nthu.edu.tw, borrissu@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw

***Abstract*—Every sound event that we receive and produce everyday carry certain emotional cues. Recently, developing computational methods to recognize induced emotion in movies using content-based modeling is gaining more attention. Most of the existing works treat this as a task of multimodal audio-visual modeling; while these approaches are promising, this type of holistic modeling underestimates the impact of various semantically meaningful events designed in movies. In specifics, acoustic sound semantics such as human sounds in movies can significantly direct the viewer's attention to emotional content in movies. This work explores the use of cross-modal attention mechanism in modeling how the verbal and non-verbal human sound semantics affect induced valence jointly with conventional audio-visual content-based modeling. Our proposed method integrates both self and cross-modal attention into a feature-based transformer (Fea-TF CSMA) where it obtains a 49.74% accuracy on seven class valence classification on the COGNIMUSE movie dataset. Further analysis reveals insights about the effect of human verbal and non-verbal acoustic sound semantics on induced valence.**

***Index Terms*—sound event detection, induced emotion, cross-modal attention, transformer**

I. Introduction

Affective analysis of movies for understanding what emotional state has been evoked in viewers is a problem of interest for multiple research communities, e.g., psychology, film-making, multimedia, etc. Computational methods for emotion modeling on multimedia data further enable applications in personalized recommendations, automated ranking systems, indexing, summarization and categorization [1]–[3]. Modeling audio-visual signals is the most prevalent strategy in developing automatic content-based affect analysis [4]. Previous works mainly focused on using audio-visual descriptors, e.g., facial expressions and audio descriptors [5], bi-modal features [6], different high-level affective audio-visual semantics features [7] and are fused in early or late fusion manner [8], [9] for predicting the movie affective responses. There are some latest studies which performs different self-attention [10] and cross-modal attention [11] mechanisms to consider the relationship among the modalities for emotion recognition; While they have shown promising accuracy, most if not all of these works considered emotion recognition as a holistic audio-video modeling task. However, a movie is orchestrated by delicately netting together relevant pieces of audio-visual elements (e.g., audio-visual scenes, speech narratives, actors acting, background music, and so on), prior works often overlooked the complex relationship between various semantic elements, such as sound events, in movies.

Moreover, it is well known that from a psychophysiological point of view, acoustic sound events can elicit or induce a series of unconscious affective responses [12]. In fact, these events are the critical composition factors in constructing a movie when intending to elicit emotion from the viewers [13]. Movie shapes the audience experiences through speech, music and sound to elicit the desired emotions; human vocal emotions in movie scenes play a critical role in directing visual attention to emotionally relevant events [14]. Essentially, these acoustic human sound events in movies, such as speech-related verbal (male speech, female speech, singing) and non-linguistic non-verbal (laughter, shouting, cry) sounds carry important cues to communicate and elicit emotion. Here, we hypothesise that focusing on acoustic human sound allows us to build more robust and powerful emotion recognition system. Our goal in this work is to integrate human vocal sound semantics together with holistic content-based audio-visual descriptors to improve induced valence states recognition. Unlike previous works where aspects of human vocal events in movies are captured partially and implicitly in the extracted features, we argue that these informative human acoustic sound semantics should be distinctively modeled to further improve the emotion recognition performance and to provide interpretable insights.

Specifically, we propose a framework that leverages the use of cross-modal attention to considers both the human vocal sound semantics and the conventional multimodal audio-visual content based modeling simultaneously. To evaluate our framework, movie annotations for both emotions and sound events are needed. There is not any appropriate movie dataset that provides both emotion and presence of sound events annotations aside from the COGNIMUSE [15]. Hence, we used the COGNIMUSE movie dataset for evaluation and achieve an accuracy of 49.74% in a seven class induced valence recognition task. Our method improves the current state-of-the-art approach by modeling acoustic sound semantics as an auxiliary yet important *modality*, and our further analysis reveals the importance of verbal and non-verbal sound semantics especially in detecting the extremes of valence states.

The rest of the paper is organized as follows: Section 2 describes the methodology. Section 3 details the experimental setups with their results and analysis, and Section 4 draws

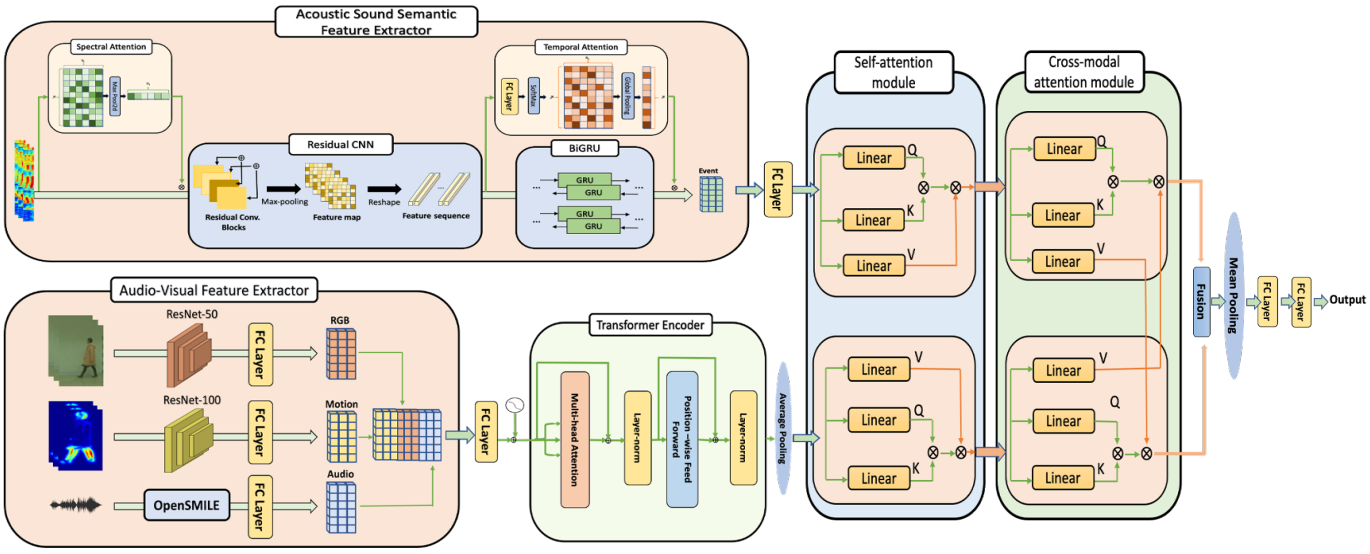

Fig. 1. Our proposed multi-modal transformer model for induced valence recognition, which includes audio-visual features and acoustic sound semantics with self and cross-modal attention mechanism

conclusion and future work.

II. Methodology

A. Dataset

For this work, the COGNIMUSE dataset [15] is used which contains half-hour continuous shots or scenes from seven high production quality Hollywood movie clips with a frame rate of 25 frames per second. They are annotated for both the emotion levels (arousal and valence ranges in [-1,1]) and the acoustic events for sounds present in the broad categories of human, mechanical, music and nature. This paper specifically focus on human sounds events in movies and emotional valence state of the viewer. Here, we consider the "experienced affective valence" annotations as our label which are the actual evoked or induced emotion when viewers immerse in the video viewing experience. To follow the settings of previous works on the same dataset, we quantize the continuous values of emotional valence into seven bins [16]. Table I lists the most occurring human sounds events with their total instances, duration, average valence and also the percent of co-occurrence rates with the extreme valence states (class 1, 2, 6, 7).

TABLE I
STATISTICS OF COGNIMUSE DATASET FOR DIFFERENT CATEGORIES OF HUMAN SOUND EVENTS WITH CO-OCCURRENCE RATES WITH THE EXTREME VALENCE STATES (CLASS 1, 2 AND 6, 7).

Category	Dur.	Inst.	Avg. Val	Co-occur(%)
ver:crowd noise	42.68	188	0.07	4.31
ver:speech male	102.39	1874	0.03	26.76
ver:speech female	55.55	1048	0.15	11.69
non-ver:shouting	8	204	-0.18	5.14
non-ver:crying	7	101	-0.33	2.53
non-ver:breathing	18.55	102	-0.37	3.68

B. Feature Extraction

1) Content-based Audio-Visual Feature Extraction: For conventional content-based audio-visual signals, we extract three different types of features (i.e., still image, optical flow and affect-related audio descriptors). To capture spatial information, the still RGB frame is used to extract the spatial features by passing the movie images to the ResNet-50 pre-trained model designed for object classification using ImageNet dataset. We take the 2048-dimensional pre-final layer features by freezing the model except for the last classification layer [7]. For motion cue information, the dense optical flow between frames is estimated by using the OpenCV optical flow algorithm and transformed into integers of [0, 255], instead of passing just two consecutive frames, we use a stack of 10 sequential frames at once to obtain a more complete motion information and feed them to the ResNet-101 pre-trained model of the ImageNet dataset classification task and extract 2048-dimensional motion features for each frame by skipping the first convolution layer and last classification layer [7]. To acquire the audio descriptors from movie audios, the OpenSMILE [17] feature extractor is used with emobase2010 configuration. The frame size and hop length are 400 ms and 40 ms respectively, this frame size corresponds to the time of stack of 10 optical flow to match the time-steps and gives a final 1582-dimensional audio feature for each frame.

2) Acoustic Sound Semantic Feature Extraction: For extracting acoustic sound semantic features, we first train a binary attention-based Convolutional Bi-directional Recurrent Neural Network (CBRNN) classifier [18] on COGNIMUSE [15] data for acoustic sound event detection (SED); here we have trained model for both verbal and non-verbal human sound events separately to detect the presence of these specific events in a movie's audio track. The log Mel-filter bank (Fbank) energies are used as the acoustic sound event detec-

TABLE II

VALENCE RECOGNITION PERFORMANCE WITH OUR PROPOSED MODEL AND OTHER COMPARED MODELS WITH VERBAL, NON-VERBAL, ALL(VERBAL+NON-VERBAL) AND WITHOUT EVENTS FEATURES. FEA-TF IS A FEATURE-BASED TRANSFORMER; TEMPO-TF IS A TEMPORAL-ATTENTION TRANSFORMER; FEA-TF CMA IS A FEATURE-BASED TRANSFORMER WITH CROSS-MODAL ATTENTION AND FEA-TF SCMA IS A FEATURE-BASED TRANSFORMER WITH SELF AND CROSS-MODAL ATTENTION.

	Conventional model								**Proposed model**			
	DNN [7]		LSTM [7]		Fea-TF		Temp-TF		Fea-TF CMA		Fea-TF SCMA	
	Acc	Acc±1	Acc	Acc±1	Acc	Acc±1	Acc	Acc±1	Acc	Acc±1	Acc	Acc±1
w/o event	43.42	90.51	37.23	89.22	**43.66**	89.64	43.11	87.97	-	-	-	-
Verbal	43.22	86.93	40.02	86.73	44.43	87.42	43.35	85.92	45.98	87.26	**47.93**	87.14
Non-verbal	43.01	85.16	39.83	84.69	43.92	86.69	42.44	84.93	44.61	86.72	**46.64**	87.58
All	43.38	87.58	40.12	86.97	44.95	88.56	43.89	86.97	47.77	87.33	**49.74**	**90.58**

tor's input and are extracted with a frame size of 400ms and a 10% overlapping window. These attentive-CBRNN human acoustic sound event detectors obtains a f1 score of 85.32%, 80.33% and 0.22%, 0.31% of an error rate for verbal-only and non-verbal-only sounds event detection respectively. In this work, we take these SED pre-trained models as a "acoustic sound-semantic" extractor to encode the audio data into human sound related acoustic event semantic features; these sound-semantic extractors are the used as an input 'modality' for our emotion recognition model.

C. Transformer Model

Motivating from the recent successful use of the transformer model [19], its encoder architecture is considered as a part of our recognition network used to learn the dependencies among the extracted content-based audio-visual features [20]. This helps us to capture the modal-specific information before cross-modeling, i.e., integration human sound semantics.

The conventional transformer encoder (TF) model contains the multi-head self-attention layer, feed-forward network. The self-attentions are calculated in parallel concerning the number of heads, and each self-attention layer initializes its query (*Q*), key (*K*), values (*V*) as weight matrices and are updated during the model training. The value weights are evaluated by the closeness or compatibility of the query with its respective key and are calculated using the following function:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

where *Q* is query, *K* is its corresponding key and d_k is the dimension of keys and queries.

$$Multihead = concat((QW_i{}^Q, KW_i{}^K, VW_i{}^V), ... \\ (QW_h{}^Q, KW_h{}^K, VW_h{}^V))W^O \quad (2)$$

where $W_i{}^Q \epsilon R^{d_m * d_k}$, $W_i{}^K \epsilon R^{d_m * d_k}$, $W_i{}^V \epsilon R^{d_m * d_v}$, $W^O \epsilon R^{hd_v * d_m}$, d_v is the dimension of the value and d_m is the model dimension.

The input of the feed-forward network is the representation obtained after applying skip residual connections on the output of multi-head self-attention.

D. Self and Cross-Modal Attention

Prior studies on emotion recognition using media data neglect the consideration of features' compatibility and suitability when it comes to fusing different modality features. Here, we design a self and cross-modal attention module to leverage the effectiveness of the cross-attention alignments to capture non-local contexts spectrally before modality fusion, i.e., human sound semantics with content-based audio-visual features. This contains two modules: self-attention module to produce their respective self-weight and cross-modal attention module to enhance the complementary discriminative power of inter and intra modality features.

In the self-attention module, firstly given input $S=\{s_1, s_2 ... s_k\}$ where $S \epsilon R^{k * d_s}$ is linearly transformed into three linear mappings query (*Q*), key (*K*) and value (*V*); $Q = SW_i{}^Q$, $K = SW_i{}^k$, $V = SW_i{}^v$ where $W_i{}^Q \epsilon R^{d_s * d_k}$, $W_i{}^k \epsilon R^{d_s * d_k}$, $W_i{}^v \epsilon R^{d_s * d_v}$ here the d_s is an input segment dimension d_v and d_k is the value and key dimensions respectively. We calculate the attention map by using *Q* and *K* which represents the feature correlation in a different position. Then, the element-wise multiplication is done with attention map and *V* to obtain the spectrally attended feature. This is computed with equation (1) separately for both the human sound semantics and the audio-visual context feature. Then, we pass these enhanced representations to the respective cross-modal attention module, here, computing attention maps are done in the same way as the self-attention module, but these energies are multiplied to other modals' *V* map to obtain a cross-modal alignment.

$$F'_{ass} = Softmax(\frac{Q_{ass} K_{ass}^T}{\sqrt{d_{kass}}})V_{avc} \quad (3)$$

$$F'_{avc} = Softmax(\frac{Q_{avc} K_{avc}^T}{\sqrt{d_{kavc}}})V_{ass} \quad (4)$$

where F'_{ass}, F'_{avc} is enhanced feature for human sound semantics and audio-visual context respectively, Q'_{ass}, K'_{ass}, V'_{ass} are query, key and value maps for human acoustic sound semantics and Q'_{avc}, K'_{avc}, V'_{avc} are query, key and, value feature for audio-visual context and d_{kass}, d_{kavc} are dimensions of respective keys and queries.

The final multi-modal representation is obtained by fusing these enhanced features followed by mean-pooling.

E. Evaluation Metrics

The above-mentioned models are evaluated based on leave-one-movie-out cross-validation and the results are presented as the average across all movie clips. As here we are dealing with long movie clips; we consider the most commonly used sound event-based metrics [21] to evaluate the performance of acoustic sound event detector. It considers the output that has a temporal position overlapping with the temporal position of an sound event with the same label in the reference and illustrate better the ability to correctly locate and label longer blocks of audio. So in this work, we integrate these human sound semantic cues in frame-wise manner with audio-visual context features and calculate frame-wise classification accuracy for predicting affective response of the viewer. Further, to give some liberty in prediction, we also estimate the classification accuracy ±1 (i.e., if the neighbouring frames are adjoined to the real class, e.g. if the real class is 2 but the model prediction is 1 or 3 then are also consider as correct) for evaluating emotion recognition on large numbers of movie frames [7].

III. Experiment and Analysis

The Adam optimizer is used in all the experiments with the learning rate and decaying factor of 0.005. The systems are trained by using back-propagation with cross-entropy loss function, and the network is trained for a maximum of 100 epochs using a batch size of 128 with early stopping. We use 64 attention dimensions for the self and cross-modal attention. For the acoustic event detection, we use the CBRNN model with parameter setting same as the previous paper [18].

A. Model Performance Comparisons

Our proposed architecture explicitly models the human verbal, non-verbal sound event semantics with content-based audio-visual signals and a feature-based transformer encoder (self-attention mechanism is applied on the extracted audio-visual features) is used to capture spectral dependency among audio-visual descriptors. Then semantically cued human sound features from acoustic sound event detectors and audio-visual context features are fed to the self and cross-modal attention module for cross-modal alignment followed by fusion. The final representation obtain after fusing and mean-pooling is passed to two fully connected layers for predicting seven-category induced valence levels.

Complete induced valence recognition results are shown in Table II includes Acc and Acc±1 of prior works and our proposed method. Here, we have re-implemented both the DNN [7] and LSTM [7] multi-model structure and later integrate the verbal and non-verbal human sound semantics explicitly to check the effect on the prediction accuracy; we consider this as the baseline for our work which is shown in the comparison Table II. From Table II, we can see that our proposed model Fea-TF SCMA (with the inclusion of acoustic human sound semantics) improves the valence classification accuracy by 6.32%, 4.51% and 3.22%, on integrating both (verbal and non-verbal), verbal and non-verbal sound semantics respectively, over the previous best work [7] (under this

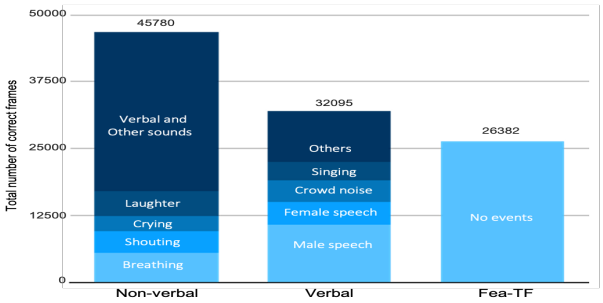

Fig. 2. The count of correct predicted frames in extreme valence (class 1, 2, 6, 7) with % of co-occur events in verbal, non-verbal and Fea-TF (w/o sound semantic inclusion) model.

setup) with 43.42% for seven category discrete-classification with frame-wise affective response prediction on the same database. Other experimented models are shown in Table II also have significant improvement integrating human event semantics, here, Fea-TF CMA (only cross-modal attention) and Fea-TF (no self or cross-modal attention) also have an increase of 4.11%, 1.29% with integrating all vocal human sound semantics, 2.32%, 0.77% for including verbal-only event semantics and 0.95%, 0.26% with non-verbal-only event semantics as compare to conventional Fea-TF (43.66% with no sound semantic inclusion). Here we also perform the traditional transformer Temp-TF (the Q calculate from previous sub layer) but the Fea-TF versions gives better prediction for induced movie emotions. To check the performance differences are statistically significant of our proposed (with sound semantic inclusion) and compared model [7] we run the T-test on the performances and get the p-value of 0.0004, 0.0034, 0.0010 for All (both verbal and non-verbal), verbal, non-verbal respectively. These p-values (all are less than 0.05) shows that the proposed models' performances differences are statistically significant.

B. Analysis of Acoustic Sound Semantic with Induced Valence

To analyse the effect of verbal and non-verbal sounds individually on induced emotional valence, we compute the accuracy recall in each emotion level and find that the human sounds in movies affect more on the extreme (class 1, 2, 6 and 7) valence with an average increase in recall of 7.39% with non-verbal and 4.67% with verbal events over Fea-TF model (without including event modality).

We also plot the total correct predicted frames for Fea-TF(w/o event), verbal, non-verbal human event semantics with the most co-occurred events. Figure 2 depicts that the inclusion of non-verbal sounds event semantics are more effective in the extreme valence than verbal (45780 correct frames vs. 32095 correct frames); for example, breathing (11.78%), crying (9.11%), shouting (6.18%) contributes more to the correct prediction. This may be due to the psychological fact that these sounds have higher temporal precedence over speech [14].

The average valence statistic in Table I in reference with the human verbal and non-verbal category events shows that the non-verbal human sound events like "crying", "breathing"

overall tend more toward the negative valence but in the case of verbal events, the event semantics like "male speech", "female speech" or "crowd noise" have the average somewhere in middle of valence scale [-1, 1]. However, we do see the co-occurrence of male speech can be 'heavy-tailed', i.e., over 25% occur in the extreme cases of valence. Table I also shows distribution of human events in extreme valence states, the human sounds such as crowd noise (4.31%), shouting (5.14%), crying (2.53%), breathing (3.68%) have a high percent of co-occurrence instead of having fewer instances in movie, this provides some insights about the association between the particular types of acoustic sound semantics and significant impact on induced valence in audiences.

IV. CONCLUSION

To enhance the recognition performance of movie-induced emotion recognition, we model the human sound semantic features in movies with conventional content-based audio-visual representations. Our trained feature-based transformer model with self and cross-modal attention (Fea-TF SCMA) performs better (49.74%) overall as compared to other models applied in the same dataset with similar emotion recognition setup. We demonstrate that the human verbal and non-verbal sound semantic features fuse well with content-based audio-visual features and indeed play an important role in predicting the evoked pleasant/non-pleasant emotional states in viewers while watching the movies. Also, our analysis shows that the different verbal and non-verbal human acoustic event semantics can have significant contributions to the extreme valence especially those of non-verbal events. For future work, we will investigate the other type of acoustic sound semantics beyond human sounds and various types of induced emotion states to better understand how the design of an *affective acoustic scene* would trigger the intended emotional state of a viewer.

REFERENCES

[1] Alan Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized tv," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 90–100, 2006.

[2] Sutjipto Arifin and Peter YK Cheung, "Affective level video segmentation by utilizing the pleasure-arousal-dominance information," *IEEE Transactions on Multimedia*, vol. 10, no. 7, pp. 1325–1341, 2008.

[3] Michal Muszynski, Leimin Tian, Catherine Lai, Johanna D. Moore, Theodoros Kostoulas, Patrizia Lombardo, Thierry Pun, and Guillaume Chanel, "Recognizing induced emotions of movie audiences from multimodal information," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 36–52, 2021.

[4] Wenzhong Guo, Jianwen Wang, and Shiping Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.

[5] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[6] Ruchir Srivastava, Sujoy Roy, Shuicheng Yan, and Terence Sim, "Multi-actor emotion recognition in movies using a bimodal approach," in *International Conference on Multimedia Modeling*. Springer, 2011, pp. 465–475.

[7] Ha Thi Phuong Thao, Dorien Herremans, and Gemma Roig, "Multimodal deep models for predicting affective responses evoked by movies.," in *ICCV Workshops*, 2019, pp. 1618–1627.

[8] Sarath Sivaprasad, Tanmayee Joshi, Rishabh Agrawal, and Niranjan Pedanekar, "Multimodal continuous prediction of emotions in movies using long short-term memory networks," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 2018, pp. 413–419.

[9] Yun Yi and Hanli Wang, "Multi-modal learning for affective content analysis in movies," *Multimedia Tools and Applications*, vol. 78, no. 10, pp. 13331–13350, 2019.

[10] Ha Thi Phuong Thao, BT Balamurali, Gemma Roig, and Dorien Herremans, "Attendaffectnet–emotion prediction of movie viewers using multimodal fusion with self-attention," *Sensors*, vol. 21, no. 24, pp. 8356, 2021.

[11] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu, "Multi-modality cross attention network for image and sentence matching. in 2020 ieee," in *CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2020, pp. 13–19.

[12] Mercede Erfanian, Andrew J Mitchell, Jian Kang, and Francesco Aletta, "The psychophysiological implications of soundscape: A systematic review of empirical literature and a research agenda," *International journal of environmental research and public health*, vol. 16, no. 19, pp. 3533, 2019.

[13] Stuart Cunningham, Harrison Ridley, Jonathan Weinel, and Richard Picking, "Audio emotion recognition using machine learning to support sound design," in *Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound*, 2019, pp. 116–123.

[14] MD Pell, K Rothermich, P Liu, S Paulmann, S Sethi, and S Rigoulot, "Preferential decoding of emotion from human non-linguistic vocalizations versus speech prosody," *Biological psychology*, vol. 111, pp. 14–25, 2015.

[15] Athanasia Zlatintsi, Petros Koutras, Georgios Evangelopoulos, Nikolaos Malandrakis, Niki Efthymiou, Katerina Pastra, Alexandros Potamianos, and Petros Maragos, "Cognimuse: A multimodal video database annotated with saliency, events, semantics and emotion with application to summarization," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, pp. 1–24, 2017.

[16] Nikos Malandrakis, Alexandros Potamianos, Georgios Evangelopoulos, and Athanasia Zlatintsi, "A supervised approach to movie emotion tracking," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 2376–2379.

[17] Florian Eyben, *Real-time speech and music classification by large audio feature space extraction*, Springer, 2015.

[18] Shreya G Upadhyay, Bo-Hao Su, and Chi-Chun Lee, "Attentive convolutional recurrent neural network using phoneme-level acoustic representation for rare sound event detection," *Proc. Interspeech 2020*, pp. 3102–3106, 2020.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2017, NIPS'17, p. 6000–6010, Curran Associates Inc.

[20] Ha Thi Phuong Thao, "Deep neural networks for predicting affective responses from movies," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4743–4747.

[21] Diego De Benito-Gorrón, Daniel Ramos, and Doroteo T Toledano, "A multi-resolution crnn-based approach for semi-supervised sound event detection in dcase 2020 challenge," *IEEE Access*, vol. 9, pp. 89029–89042, 2021.