INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Predicting Group-Level Skin Attention to Short Movies from Audio-Based LSTM-Mixture of Experts Models

Ricardo Kleinlein[1], *Cristina Luna Jiménez*[1], *Juan Manuel Montero*[1], *Zoraida Callejas*[2], *Fernando Fernández-Martínez*[1]

[1] Information Processing and Telecommunications Center, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid, 28040, Madrid, Spain
[2] Department of Languages and Computer Systems, University of Granada, 18071, Granada, Spain

ricardo.kleinlein@upm.es, cristina.lunaj@upm.es, juanmanuel.montero@upm.es, zoraida@ugr.es, fernando.fernandezm@upm.es

Abstract

Electrodermal activity (EDA) is a psychophysiological indicator that can be considered a somatic marker of the emotional and attentional reaction of subjects towards stimuli like audiovisual content. EDA measurements are not biased by the cognitive process of giving an opinion or a score to characterize the subjective perception, and group-level EDA recordings integrate the reaction of an audience, thus reducing the signal noise. This paper contributes to the field of audience's attention prediction to video content, extending previous novel work on the use of EDA as ground truth for prediction algorithms. Videos are segmented into shorter clips attending to the audience's increasing or decreasing attention, and we process videos' audio waveform to extract meaningful aural embeddings from a VGGish model pretrained on the Audioset database. While previous similar work on attention level prediction using only audio accomplished 69.83% accuracy, we propose a Mixture of Experts approach to train a binary classifier that outperforms the main existing state-of-the-art approaches predicting increasing and decreasing attention levels with 81.76% accuracy. These results confirm the usefulness of providing acoustic features with a semantic significance, and the convenience of considering experts over partitions of the dataset in order to predict group-level attention from audio.

Index Terms: Electrodermal activity, attention prediction, affective video content analysis, recurrent neural network, mixture of experts

1. Introduction and related work

The question on how humans react towards stimuli has been addressed from many and varied perspectives such as psychology, medicine or neuroscience. Attention was defined at the end of the XIX century by Wilhelm Wundt and William James as "taking possession by the mind, in clear and vivid form, of one of what seem several simultaneously possible objects or trains of thought" [1]. Research on psychophysiology has led to the creation of several techniques to measure the activity of the central nervous system (CNS), making possible the study of mind expressions, including attention. Some of such methods are functional magnetic resonance imaging (fMRI) and positron emission tomography (PET), which have been both used in research on emotions [2], yielding in some cases to revolutionary findings [3]. Another method that has been employed in this regard is galvanic skin response (GSR), often also referred to as electrodermal activity (EDA) [4]. EDA, as opposed to the previously mentioned methods, measures the autonomic nervous system (ANS) peripheric responses using electricity [5]. It refers to the variation of the electrical properties of the skin due to the secretion of sweat by eccrine glands, under the control of the sympathetic autonomous system (ANS) which is, in turn, influenced by the CNS. These electrical variations are considered somatic markers of attentional and emotional expressions not biased by the cognitive process of giving an opinion to characterize the subjective perception [4, 6, 7].

Within this context, a neuromarketing technology based on EDA, Sociograph [8, 9], aims at registering the attentional and emotional response of groups of people when they are presented some kind of stimuli, for instance, during the screening of an audiovisual composition. The EDA of up to 128 subjects can be registered at the same time by the Sociograph technology, which encompasses both hardware and software. The outcome is the aggregation of the synchronized signals from all the subjects of a screening session. This output signaling can be referred to as group-level electrodermal activity (EDAg). An example of EDAg recordings is shown in Figure 1.

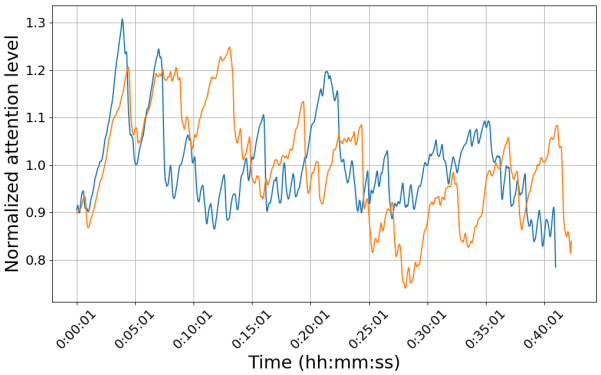

Figure 1: *Two examples of EDAg recordings during two screening sessions, after isolating the signal corresponding to the attention level.*

The seminal work by Lang *et al.* proved that electrodermal activity is useful to characterize emotions reported by subjects when watching pictures [10]. In Fleureau *et al.* a similar experiment applied to videos was carried out [11]. Content-based descriptors and EDA were combined in Soleymani *et al.* to model spectators' arousal and valence [12]. Datasets such as DEAP [13], LIRIS-ACCEDE [14], RECOLA [15] and MAHNOB-HCI [16] collect several physiological signals, including EDA recordings, further supporting the use of electro-

http://dx.doi.org/10.21437/Interspeech.2019-2799

dermal activity as an important indicator of the attentional and emotional state of a person. Nevertheless, most authors have typically focused in the prediction of the EEG signal, while the modelling of EDA responses using content-based audiovisual features has been tackled less often [17].

In Hernández-García *et al.* [18], a linear regression based on a set of 31 low-level visual features was used to predict the emotion and attention levels from the EDA recordings of an audience watching the awarded spots of the 2002 Cannes International Advertising Festival. Similarly, García-Faura *et al.* considered a logistic regression classifier based on both audio and visual low-level descriptors in their analysis [19]. Nonetheless, the aural features employed were related exclusively to music-oriented acoustic properties. It is clear to us that the narrative structure and the semantics of a movie largely influence a viewer's attention [20]. We go further in this line of research by processing the audio signal in a way such that our predictive algorithms are fed with semantic information about the movie along time.

Such acoustic features will have the form of complex non-linear relationships between input variables, unfit for linear machine learning methods. Therefore the vast majority of algorithms in speech and audio technologies nowadays are based on deep neural networks, due to the intrinsic complexity of audio signaling [21, 22, 23]. Ensemble models can be thought as a set of submodels in which the global output is given by the weighting of those submodels' [24]. The models do not need to meet any particular requirement, so a combination of neural networks is perfectly possible. The submodels do not access the same data, but rather specialize in subsets of the data distribution. This enables the creation of experts. This mixture of experts (MEX) can be either cooperative (when the global output comes as a linear combination of the experts' outputs) or competitive, when just one expert evaluates the input and is fully responsible for the global response [25]. Many applications have benefited from this approach since it allows a model to characterize data distributions that can be better understood in terms of more than one local regime within the data distribution [26, 27].

We aim at building MEX models that make an effective use of aural vectors with rich semantic information to model the attention reaction they evoke in the audience. We are not analyzing the psychophysiological reactions to videos but we take these reactions as ground truth to label video clips accordingly.

2. Short movies dataset

In order to collect ground truth data, EDA was measured on participants while they watched a concatenation of videos selected from 136 short films from the Jameson Notodofilmfest Short Film Festival 2015. The age of the annotators ranged from 17 to 60 years old with an average of 23.11 and 46.4% of male participants. Twelve screening sessions were held, each with a mean number of 22.5 attendees from a total of 270 participants. The EDA was recorded at 1 Hz on each subject by means of the Sociograph device. Each short film was segmented into shorter clips attending to the attention level slope, separating those scenes that resulted in increasing attention from the ones that made the audience lose interest. For each clip, Sociograph integrates the measurements from all participants and outputs separate signals for attention and emotion of the group along the video with a 95% confidence. A complete description on the process followed to compute attention indicators from the raw EDAg signal can be found in [19].

Instead of using audio raw waveforms we convert them into semantically rich vector embeddings each covering 1 second of audio, far more related to the narrative structure of the movie clip than other typically used representations such as MFCC features. The process to compute embeddings from audio waveforms is that of [23], which spans two steps which respective outputs are displayed in Figure 2. First of all we generate log-mel spectrograms from the audio waveforms, and afterwards we feed these spectrograms to a VGGish-like audio classification model [23] that outputs a dense layer of embeddings. VGGish is a variant of the original configuration A with 11 weight layers of a VGG model [28], but whose last layer is a 128-wide fully connected layer. The log-mel features are computed at a sample rate of 16kHz, one channel, and using magnitudes of the Short-Time Fourier Transform with window size of 25ms, window hop of 10 ms and a periodic Hann window, mapping the spectrogram to 64 mel bins covering the range between 125 and 7500 Hz. These features are then framed into non-overlapping instances of 0.96 seconds of length, so each sample covers 64 mel bands and 96 frames of 10 ms of duration each. The last layer of VGGish acts as a compact embedding layer and conforms the embedding vectors we consider throughout this study as inputs for the predictive models after we apply a 8-bit quantization on them.

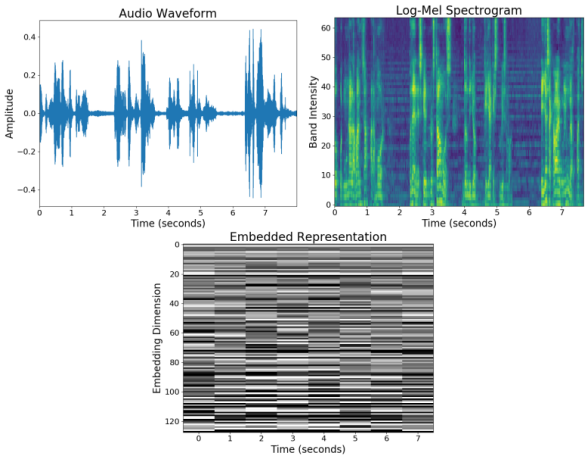

Figure 2: *Audio representations at the 3 different stages of the audio processing pipeline. The embeddings implicitly attach a dense semantic meaning, as opposed to the waveform and the log-mel spectrogram.*

The VGGish model we used is the default version pre-trained on Audioset [29] and described in [23]. The Audioset database consists of about 2 million human-labeled 10-second sound clips extracted from Youtube videos [29]. The labels to these videos are drawn from an ontology of 632 possible audio event categories, specified as a hierarchical graph of event categories. This dataset covers a wide range of sounds belonging to human, animal, musical instruments and common environmental sounds. This diversity ensures the embeddings generated are rich in semantic content, effectively telling between sounds coming from different sources such as speech, music or other audio events. Having a robust embedding generation procedure instead of fixed audio categories, over as many types of audio events as possible is of utmost importance, since the audio files we consider belong to free-topic short movies, in which a high degree of overlapping between

acoustic events is expected. The vector embeddings, as well as code and annotations used in this study are publicly available in *https://github.com/ricardokleinklein/audio_attention*.

Additional information from the participants in the screening sessions is available, such as the movie genre and whether the participant liked it or not. The 10 genres present in our dataset define in broad terms the narrative structure of the film (*animation, action/adventure, science-fiction, comedy, drama, documentary, thriller, social issues, terror and other*). We assume that all clips belonging to the same movie share the same genre, and for every film at least one genre was assigned. When multiple genres were associated to a movie, only the most relevant one was taken. After watching each movie, every participant in the audience was asked to give an opinion on it, indicating their affective perception of it in terms of *like/neutral/dislike*. Therefore for every movie a distribution of scores indicating how many viewers liked, disliked or had a neutral opinion was available. Neither of the two aforementioned variables (movie genre and affective perception score) was found to have a direct correlation with the attention level but rather provide further information related to the participants' screening experience.

3. Group-level skin response prediction networks

In this section we describe the models that constitute our predictive algorithms. First we introduce the simplest model (canonical model), which is the starting point from which we train specialized experts on different partitions of the data, as will be explained in Section 3.2. We propose dynamic models that have the capacity to observe the local evolution of the acoustic events along the input signal, contrarily to previous approaches based on temporary-static models [18, 19]. Because the audio clips do not have in general a similar time length (Table 1), models that do not observe the dynamic evolution might be not well-suited for these kind of tasks.

Table 1: *Mean time length and standard deviation of video clips according to their attention tag.*

Attention Level	μ (s)	σ (s)
Increasing	41.1	42.73
Decreasing	7.56	7.22

3.1. Canonical model

Given the sequential nature of the input data, we chose a recurrent unidirectional LSTM cell layer as the main building block for our models [30]. In particular, our canonical model is made up of 3 layers of such cells, and only the last hidden state of the LSTM is passed to a final fully-connected layer. The equations defining our LSTM cells are given by:

$$i_t = \sigma\left(W_{ii}x_t + b_{ii} + W_{hi}h_{(t-1)} + b_{hi}\right) \quad (1)$$

$$f_t = \sigma\left(W_{if}x_t + b_{if} + W_{hf}h_{(t-1)} + b_{hf}\right) \quad (2)$$

$$g_t = \tanh\left(W_{ig}x_t + b_{ig} + W_{hg}h_{(t-1)} + b_{hg}\right) \quad (3)$$

$$o_t = \sigma\left(W_{io}x_t + b_{io} + W_{ho}h_{(t-1)} + b_{ho}\right) \quad (4)$$

$$c_t = f_t c_{(t-1)} + i_t g_t \quad (5)$$

$$h_t = o_t \tanh\left(c_t\right) \quad (6)$$

Table 2: *Summary of the data partitions considered.*

MEX	Partition — Expert Model		# Clips	Time Length (h)	Majority class (%)
Perception	2Exps	Like	207	1.75	53.40
		Dislike	330	2.26	61.21
	Neutral	Like	207	1.75	53.40
		Neutral	140	0.97	61.33
		Dislike	190	1.29	61.11
	Neutral-Strict	Like	207	1.75	53.40
		Neutral	235	1.60	61.02
		Dislike	95	0.65	61.70
Genre	Comedy		123	0.74	60.38
	Drama		192	1.40	55.86
	Other		222	1.86	59.62

The input at each time, x_t for $t \in [0, 1, 2, ..., T]$, with T being the length in seconds of the sequence, is the 128-D embedding vector computed with VGGish.

3.2. Mixture of Experts (MEX)

The complementary information of the clips such as the movie genre or the affective perception rating can be used to divide the set of clips into smaller, more specialized sets of data, and train attention prediction experts accordingly to those partitions. Depending on the information used (either perception or genre) we can identify different partitions. In the case of perception, because of the cognitive process of giving an opinion about a movie, a broad variety of responses from the audience was available, so we have tried several different partitions that have been chosen following the decision tree schemed in Figure 3. When focusing on the genre, the amount of time available in the clips belonging to each genre was the criterion used to define the data partitions, since not all movies have the same length.

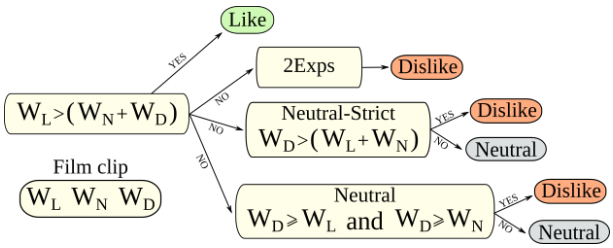

Figure 3: *Data partition rules for attention prediction MEXs based on affective perception. W_L, W_D and W_N denote the proportion of annotators who liked, disliked and had a neutral opinion on the film, respectively.*

In all cases, the experts are trained following a domain adaptation scheme: starting from the best canonical model obtained, the models are fine-tuned on the specific segments of the data we want each of them to specialize on. This way we ensure experts perform at least as well as the canonical model they are trained from, never worse. Table 2 shows that regardless of the partition strategy used, each expert gets to see sets of samples fairly comparable to the rest of experts. These partitions suggest convenient starting points for attention prediction MEX approaches based on different criteria, for neither the affective perception nor the genre seem to be directly correlated to attention.

4. Experimental results and discussion

We adopted a stratified 10-fold cross-validation analysis as our standard experimental setup. In all cases the training was carried out using an Adam optimizer [31], with learning rate of

Table 3: *Accuracy rates obtained when evaluating different families of algorithms over the same dataset of short movies. Mixture of Experts approaches perform better than any low-level features-based or canonical models.*

Approach	Feat. Type / Model	Feat. Dim.	Accuracy (%)(σ)
Baseline [19] (clip-level functionals)	LR (Only aural)	50	69.83 ± 3.88
	LR (only visual)	10	74.98 ± 3.66
	LR (visual + aural)	60 (50 + 10)	79.59 ± 3.41
Canonical (frame-level embeddings)	VGGish + LSTM	64	78.08 ± 3.50
		128	78.65 ± 3.47
		256	78.46 ± 3.48
MEX (frame-level embeddings)	VGGish + LSTM (2Exps)	128	79.88 ± 3.39
	VGGish + LSTM (Neutral)	128	81.76 ± 3.27
	VGGish + LSTM (Neutral-Strict)	128	81.62 ± 3.28
	VGGish + LSTM (Genre)	128	80.10 ± 3.38

Table 4: *Ablation study in attention prediction between MEXs and canonical models for all the data partitions studied.*

MEX	Partition — Data split		Model Attention Accuracy			
			Like		Dislike	Canonical
Perception	2Exps	Global	78.03		78.03	78.59
		Like	**82.21**		77.31	80.19
		Dislike	75.46		**78.49**	77.58
			Like	Neutral	Dislike	Canonical
	Neutral	Global	79.14	79.14	78.22	78.77
		Like	**83.59**	80.21	77.36	81.17
		Neutral	78.57	**80.71**	76.43	78.57
		Dislike	75.26	76.84	**80.53**	76.32
	Neutral-Strict	Global	79.22	78.87	79.00	79.03
		Like	**82.17**	76.45	79.24	79.31
		Neutral	76.68	**82.19**	78.77	79.64
		Dislike	79.00	75.89	**79.00**	76.89

		Comedy	Drama	Other	Canonical
Genre	Global	76.52	77.99	78.37	78.2
	Comedy	**77.44**	73.4	73.4	73.4
	Drama	76.42	**78.53**	76.42	77.47
	Other	76.15	80.2	**82.92**	81.56

5×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and exponential learning rate decay of 0.95, one audio file at a time. This setup was kept the same throughout this study.

Table 3 presents the accuracy rates achieved with both low-level functionals, canonical and MEX models evaluated over the short movies dataset described in section 2. Canonical networks outperformed a logistic regression model (LR) trained on low-level features of a single modality. Nonetheless, MEX approaches yielded the highest accuracy rates in all cases, with better results than canonical models or LR trained on both audio and image modalities. In particular, models that exploited partitions of the data based on the affective perception of the film (*Neutral*, *Neutral-Strict*) were the ones that benefited the most from this strategy.

Table 4 shows an ablation study of both the canonical model and the experts for all the data partitions considered in this study. It can be noticed that in all cases, the individual expert models perform better than the canonical model over their respective data partition. This is because the learners concentrate on extracting relevant features for each of the partitions considered, finding more appropriate representations of the data than the canonical model. In fact, the more balanced the partitions are, the higher the accuracy rate achieved by the corresponding MEX model is.

These results reinforce the idea that a wise partitioning of the data can offer a better opportunity to fully exploit the information contained in the data set. Because of this data segmentation, a delicate equilibrium between the number of partitions we define and the amount of specialization we can expect from the experts must be taken into account.

5. Conclusions

While electrodermal activity (EDA) has been used for long time in psychology and medicine, and more recently in neuroscience and neuromarketing, as a way of measuring the reaction of people towards stimuli, little research has been devoted by the affective computing field to the stimuli side relying on EDA as ground truth for attention assessment. EDA might become a popular measure of ground truth annotation for video content analysis in future research, since it is a relatively straightforward and non-expensive method for capturing the emotional states of the human mind. We segment videos in shorter clips, which are then labeled with regard to the attention slope (increasing vs. decreasing). Then, we compute a set of sequences of semantically-dense vector embeddings describing the audio from the videos to obtain an aural representation more related to the narrative structure of the clips, which is fundamental in driving audience's attention. We have released this dataset so further research can be carried out. This work reveals that audio alone is useful to model viewers' reaction to audiovisual stimuli in terms of EDA.

The analysis of these sequences by recurrent neural networks has outperformed our previous attempts using only aural features such as logistic regression, validating the use of deep learning-generated aural feature extraction. It is remarkable that our exclusively audio-based MEX models achieve even greater accuracy than models based on descriptors of both audio and image. This result points out that the extraction of features via deep learning processing is a successful approach that is helping in understanding the development of the narrative within the scene that has to do with attention prediction. Furthermore, the difference in the average time length between the clip classes considered (increasing/decreasing attention) suggests that our prediction could benefit from turning the problem into a regression task considering the whole film instead of classifying short clips labeled attending to the attention level.

Exploiting the possible segmentation of the original dataset in terms of the additional information available has proved to be a succesful approach, particularly when the subjective information of perception via score was considered. However, a delicate data partitioning must be carried out, given the diversity in the samples' content, which adds complexity to the task of predicting the attention level of a movie clip. Furthermore, since all the videos in the dataset are uploaded to Youtube, as a future work we could compare how audience reaction in terms of EDA correlates with the conscious opinion given by the online community, following a similar procedure to the one in Fernández-Martínez *et al.*[32]. Different MEX systems could be proposed attending to the descriptors available from Youtube to create more accurate models of attention. Moreover, a larger set of annotated data would shed light on different MEX selection rules.

6. Acknowledgements

The work leading to these results has been supported by the Spanish Ministry of Economy, Industry and Competitiveness through the ESITUR (MINECO, RTC-2016-5305-7), CAVIAR (MINECO, TEC2017-84593-C2-1-R), and AMIC(MINECO, TIN2017-85854-C4-4-R) projects (AEI/FEDER, UE).

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for part of this research. We also appreciate the collaboration of Sociograph Neuromarketing in providing the data set and annotations.

7. References

[1] W. James, *The pinciples of psychology*. New York; Dover, 1890.

[2] K. Phan, T. Wager, S. F. Taylor, and I. Liberzon, "Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in pet and fmri," *NeuroImage*, vol. 16, no. 2, pp. 331 – 348, 2002.

[3] R. Cabeza and L. Nyberg, "Imaging cognition ii: An empirical review of 275 pet and fmri studies," *J. Cognitive Neuroscience*, vol. 12, no. 1, pp. 1–47, Jan. 2000.

[4] M. Dawson, A. Schell, and D. Filion, "The electrodermal system," in *Handbook of Psychophysiology*, 01 2000, pp. 200–223.

[5] C. Féré, *Note sur les modifications de la résistence électrique sous l'influence des excitations sensorielles et des émotions*. CR Soc. Biol, 1888, vol. 5.

[6] W. Boucsein, *Electrodermal activity*. Springer Science & Business Media, 2012.

[7] H. Sequeira, P. Hot, L. Silvert, and S. Delplanque, "Electrical autonomic correlates of emotion," *International Journal of Psychophysiology*, vol. 71, pp. 50–56, 01 2009.

[8] J. L. Martínez and E. Garrido, *Sistema para la medición de reacciones emocionales en grupos sociales*. 168 928, 2003, vol. 2.

[9] M. Aiger, M. Palacín, and J.-M. Cornejo, "Electrodermal signal by sociograph: methodology to measure the group activity," *Revista de Psicología Social*, vol. 28, no. 3, pp. 333–347, 2013.

[10] P. J. Lang, M. Greenwald, M. M. Bradley, and A. Hamm, "Looking at pictures: Affective, facial, visceral, and behavioral reactions," *Psychophysiology*, vol. 30, pp. 261–73, 06 1993.

[11] J. Fleureau, P. Guillotel, and I. Orlac, "Affective benchmarking of movies based on the physiological responses of a real audience," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Sep. 2013, pp. 73–78.

[12] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses," in *2008 Tenth IEEE International Symposium on Multimedia*, Dec 2008, pp. 228–235.

[13] S. Koelstra, C. Muhl, M. Soleymani, J. 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.

[14] Y. Baveye, E. Dellandra, C. Chamaret, and L. Chen, "Liris-accede: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, Jan 2015.

[15] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, April 2013, pp. 1–8.

[16] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, Jan 2012.

[17] D. Ayata, Y. Yaslan, and M. Kamaak, "Emotion recognition via random forest and galvanic skin response: Comparison of time based feature sets, window sizes and wavelet approaches," in *2016 Medical Technologies National Congress (TIPTEKNO)*, Oct 2016, pp. 1–4.

[18] A. Hernández-García, F. Fernández-Martínez, and F. Díaz-de María, "Emotion and attention: Predicting electrodermal activity through video visual descriptors," in *Proceedings of the International Conference on Web Intelligence*, ser. WI '17. New York, NY, USA: ACM, 2017, pp. 914–923.

[19] A. García-Faura, A. Hernandez-Garcia, F. Fernández-Martínez, F. Díaz-de María, and R. San-Segundo, "Emotion and attention: Audiovisual models for group-level skin response recognition in short movies," *Web Intelligence and Agent Systems*, vol. 17, pp. 29–40, 02 2019.

[20] E. Tan, *Emotion and Structure of Narrative Film: Film as an Emotion Machine*. Lawrence Erlbaum Associates, Inc., 1996.

[21] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[22] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Interspeech 2017*, 08 2017, pp. 4006–4010.

[23] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[24] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, Mar. 1991.

[25] D. B. Ferrari and A. Z. Milioni, "Choices and pitfalls concerning mixture-of-experts modeling," *Pesquisa Operacional*, vol. 31, pp. 95 – 111, 04 2011.

[26] D. Eigen, M. Ranzato, and I. Sutskever, "Learning factored representations in a deep mixture of experts," *CoRR*, vol. abs/1312.4314, 2013.

[27] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *CoRR*, vol. abs/1701.06538, 2017.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[29] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[32] F. Fernández-Martínez, A. Hernández-García, A. Gallardo-Antolín, and F. Díaz-de María, *Combining audio-visual features for viewers' perception classification of Youtube car commercials*. International Speech Communication Association., 2014.