AN AUDIO-SALIENCY MASKING TRANSFORMER FOR AUDIO EMOTION CLASSIFICATION IN MOVIES

Ya-Tse Wu, Jeng-Lin Li, Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University, Taiwan

ABSTRACT

The process of perception to affective response of humans is gated by a bottom-up saliency mechanism at the sensory level. In specifics, auditory saliency emphasizes audio segments that need to be attended to cognitively appraise and experience emotion. In this work, inspired by this mechanism, we propose an end-to-end feature masking network for audio emotion recognition in movies. Our proposed Audio-Saliency Masking Transformer (ASTM) adjusts feature embedding using two learnable masks; one of them cross-refers to an auditory saliency map, and the other one is through self-reference. By joint training for front-end mask gating and the transformer as the back-end emotion classifier, we achieve three-class UARs of 46.26%, 49.03%, 53.49% and 53.51% on experienced arousal, experienced valence, intended arousal, and intended valence, respectively. We further analyze which acoustic feature categories that our saliency mask attends to the most.

Index Terms— emotion recognition, auditory saliency, affective multimedia, transformer

1. INTRODUCTION

Movie films contain realistic emotion elicitation needed to trigger emotional experiences for audiences. The ability to automatically recognize the audience's affective viewing experiences using content-based modeling is important for a wide range of multimedia applications such as user-centered recommendation [1] and intelligent indexing [2]. In fact, most prior research in emotion recognition has focused more on processing multiple data streams as measurements from an individual but less on processing the affective media content itself as impact on an individual. In this work, we focus on modeling movie content to recognize both induced and experienced emotion states. Specifically, we target audio tracks in movies. The audio track of a movie is not only an information-rich modality which includes a multitude of auditory components (e.g., audio events, speech, soundtrack music, and so on) but also provides large-scaled realistic media data for research.

Human's perceptual process of taking in a sensory input and triggering an internal emotional response involves three major components, i.e, sensory, cognition, and emotion [3]. The hierarchical integration from signal-level auditory input to high-level cognitive/appraisal process forms the final emotion responses [4]. There are two different attention mechanisms in this process: the bottom-up and top-down attention. The bottom-up attention is known as "saliency", which is a signal-based attention that acts as a gating filter to naturally gear our sensory attention [5]. Top-down attention, which is a cognitive and task-driven attention, is also seen as a process of gaining control for specific tasks at hand [6]. Handling the extreme complexity in performing content-based modeling for emotion recognition has benefited from the use of the top-down attention mechanism in deep networks [7].

The top-down attention mechanism, i.e., commonly used in deep learning, can be seen as task-based saliency (e.g., derived discriminatively based on emotion recognition task) but not a signal-based (bottom-up) saliency [6]. Some research has integrated bottom-up auditory saliency for speech tasks, e.g., using an auditory saliency spectral mask for noise-robust speech recognition [8], and improving cognitive load classification by pooling saliency mask over time [9]. While integrating saliency to spectral representation has been useful, few studies have utilized signal-level saliency for emotion tasks. Work by Aldeneh and Provost is one of the few that integrated region-based saliency on Mel Filterbank for speech emotion recognition [10]. Very few, if any, have modeled this tiered saliency-to-attention in content-based modeling for affective multimedia applications. The challenge is likely due to the fact that using spectrograms solely may not be sufficient for emotion modeling while the auditory saliency mask is known to derive only for the spectral representation.

Knowing that emotion processing involves a saliency to attention scheme, we hypothesize that auditory saliency helps provide a mechanism in "gating" the sensory input that helps cognitively attend to the right information. In this work, we model such a process by proposing an Audio-Saliency Masking Transformer (ASMT). ASMT is an end-to-end transformer network. The front-end is built by a trainable feature masking and the back-end is a transformer classifier. There are two learnable masks during training, one of them is an acoustics descriptor self-adjusted mask, while another is a cross-references to the auditory saliency map, i.e., adjusted auditory saliency mask. The linear combination of acoustic descriptors is enhanced by these two masks and further integrated before passing it to the transformer. The final

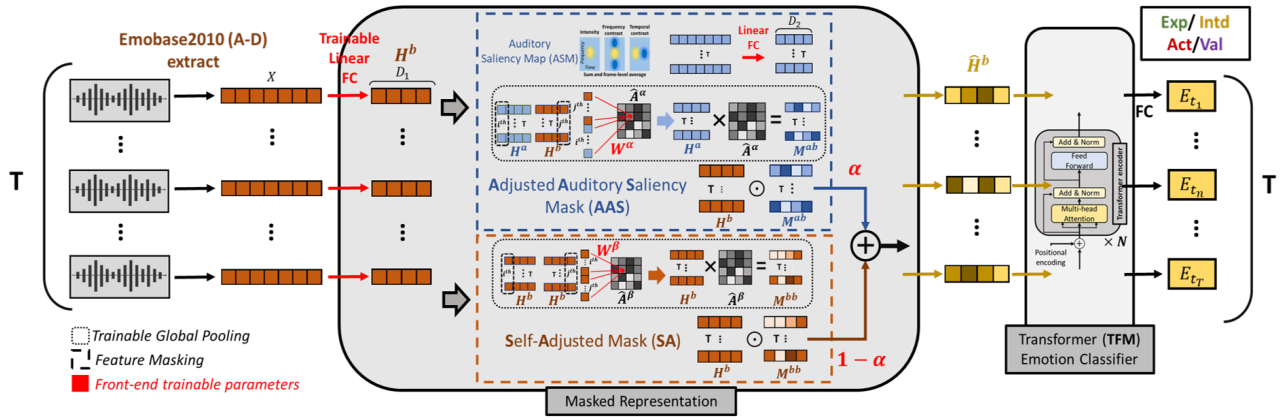

Fig. 1. Our proposed framework of ASMT. H^b is obtained by passing Emobase2010 (A-D) through the trainable linear fully connected layer (FC) with dimension D_1. Front-end representation $\hat{H}_b$ is summarized by trainable α of the self-adjusted mask (M^{bb}) and the adjusted auditory mask (M^{ab}); both masks are derived from trainable global pooling mechanism with parameters W^α and W^β, respectively. Finally, $\hat{H}_b$ passes to a back-end transformer for emotion classification.

classification result reaches state-of-the-art 46.26%, 49.03%, 53.49% and 53.51% on three classes in four different attributes: experienced arousal, experienced valence, intended arousal, intended valence, respectively.

2. RESEARCH METHODOLOGY

2.1. Dataset

COGNIMUSE database is a multi-modal movie dataset [11]. It includes twelve half-hour continuous movie clips. Emotion annotation is annotated in continuous time with arousal and valence value within the range [-1,1]. There are two different types: intended and experienced. Intended emotion indicates the response that the movie tries to evoke in the viewer, without taking into account whether it is actually successful. The experienced emotion means to rate the actual emotional experience when watching movies [12]. In this work, we recognize both intended and experienced emotion labels in twelve movie clips. Following a similar setup [13], we average every emotion annotation in an non-overlapping 5-seconds intervals of audio track. Finally, we obtain 4414 audio segments as our dataset where each sample has two different types of arousal and valence ratings.

2.1.1. Discretizing Emotion Label

The original COGNIMUSE provides frame-wise continuous emotion annotation, since our work focuses on recognizing emotion for every 5 second of audio track segments, we make it a classification problem. Since both emotions are annotated by multiple subjects, we used the average values from each subject. Seven classes discretization has been suggested to quantize continuous time emotion label, however, due to an imbalance of class distribution, we reduce seven classes to three classes (with equally-spaced intervals) indicating "high", "mid", "low" for both arousal and valence. The experienced emotions are further re-scaled via min-max normalization before applying discretization. The number of samples per class is shown in Table 1.

Table 1. Discretized emotion label distribution.

Type	Experienced		Intended	
	Act	Val	Act	Val
Low	2845	695	1801	1578
Mid	1480	2257	1428	874
High	89	1462	1185	1962

2.1.2. Acoustic Descriptors (A-D)

We extract the emobase2010 feature set, which is a 1582 dimensional feature, extracted using the OpenSMILE toolkit from each 5-seconds segment [14]. Emobase2010 feature sets, which contain acoustic descriptors (A-D), are known to capture various acoustic-prosodic properties of an audio segment. Previous works have also used emobase2010 to trace affective content in the movie [15].

2.2. Auditory Saliency Map (ASM)

Auditory saliency map (ASM) is a mask representing the degree of saliency in time-frequency domain [16]. The concept of ASM is to model the human's bottom-up auditory attention using sets of 2D Gabor filters. Assume n is the time dimension, w is the frequency dimension, $\hat{S}(n,w)$ is log magnitude spectrogram, and $G(n,w)$ is 2D Gabor filters, convolution output is $R_k(n,w) = \hat{S}_k(n,w) * G(n,w), \quad k = 0, ..., K-1$. In our work, three individual $G(n,w)$ are applied to capture intensity, frequency contrast, and temporal contrast, respectively. The 2D representations R_k are up-sampled to $\hat{R}_k$ by using interpolation [8]. After the interpolation, subtracting the center (c) with the surround (s) is applied: $F(c,s) = \hat{R}_c - \hat{R}_s$, $c = \{0, ..., K-3\}, s = \{c + \Delta\}, \ \Delta = \{1, 2\}$. The result of center-surround difference $F(c,s)$ is thresholded by zeros and normalized to local maximum. The final ASM is the average of three individual center-surround difference representations (intensity, frequency contrast, temporal contrast). We extract ASM with DFT $size = 1024$ and $K = 4$. Mean pooling is done over 5-seconds segment on ASM results to obtain a 257 dimensional features representing auditory saliency across different frequencies.

2.3. Audio-Saliency Masking Transfomer (ASMT)

The entire structure is shown in Fig. 1. Emobase2010 is first passed through a linear fully-connected layer with dimension D_1 in sets {32, 64} and a dropout rate 0.5 to reduce feature dimension. ASMT is an end-to-end architecture that takes a sequence T acoustic descriptors and output a sequence of emotion labels; the transformer is trained as a back-end, and the front-end is a composite saliency masked representations.

The composite representations used as front-end $\hat{H}^b \in \mathbb{R}^{T\times D_1}$ is obtained through a trainable parameter α that sums the two representations $H^{ab} \in \mathbb{R}^{T\times D_1}$ and $H^{bb} \in \mathbb{R}^{T\times D_1}$:

$$\hat{H}^b = \alpha H^{ab} + (1-\alpha)H^{bb} \tag{1}$$

One of them is a self-masking representation (H^{bb}) while another one is an auditory saliency masking representation (H^{ab}). Masking is defined as a learnable matrix that can be multiplied to a sequence T acoustic descriptors (A-D) segments to obtain the "masked" (enhanced) representations. We will talk about each of the masks and the strategy in learning the masking matrices below.

2.3.1. Masked Representations

Two different representations H^{ab}, H^{bb} are computed by:

$$\begin{aligned} H^{ab} &= H^b \odot M^{ab} \\ H^{bb} &= H^b \odot M^{bb} \end{aligned} \tag{2}$$

$\odot$ is the element-wise product, the masks are in dimension (T, D_1). The idea is to learn a global 2-D matrix corresponding to the feature set (ASM and A-D or A-D and A-D). It re-weights the importance of the original acoustic descriptors to derive masked representation. Two different masks are derived by:

$$\begin{aligned} M^{ab} &= H^a \hat{A}^\alpha \\ M^{bb} &= H^b \hat{A}^\beta \end{aligned} \tag{3}$$

where $\hat{A}^\alpha \in \mathbb{R}^{D_2\times D_1}$ and $\hat{A}^\beta \in \mathbb{R}^{D_1\times D_1}$ are the global 2-D matrix for deriving masks, and D_2 is the hidden dimension from another branch (ASM). M^{bb} is an acoustic descriptor self-adjusted mask (SA). M^{ab} is an adjusted auditory saliency mask (AAS), which is modulated from the ASM, integrating the prior knowledge of spectral based auditory saliency. For every value in each mask, it indicates the weight (importance) of a particular feature dimension. It refers to either other dimensions of the same features (self-adjusted) or dimensions of spectral saliency knowledge (ASM). We devise a trainable global pooling mechanism in order to derive the value of each of these masking matrices.

2.3.2. Trainable Global Pooling Mechanism

To identify the importance value of a feature sets H^b at j dimension either through self-reference (self-adjusted) or cross-reference (to ASM), we design a linear transformation W acting as a global pooling matrix that take each dimension of (i, j) in (H^a, H^b) set to a scalar to learn a value in the mask.

$$\begin{aligned} A^\alpha_{i\,j} &= W^\alpha({H^a}_T{}_{,i}||{H^b}_T{}_{,j}), \\ \hat{A}^\alpha_{i\,j} &= \frac{exp(A_{ij})}{\sum_{k=0}^{D_2-1} exp(A_{ik})} \end{aligned} \tag{4}$$

This trainable matrix $W^\alpha \in \mathbb{R}^{1\times 2T}$ is designed to learn the summarization of discriminatively-common portion between two different sets of the features then to act as a gating value to the feature sequence. The challenges in learning a mask when incorporating prior knowledge is that two feature sets may be inherently different in dimensions with distinct physical meanings (in this case, one is auditory spectrum and another one is acoustic descriptors). This trainable global pooling mechanism obtains the importance in dimension of different feature sets to derive the masks, and it further enforces the use of few parameters in W^α to aggregate the needed global pattern between feature sets. In Fig. 1, $\hat{A}^\beta$, H^b and W^β replace $\hat{A}^\alpha$, H^a and W^α in equation (4) respectively, and obtain an A-D self-reference mask M^{bb}.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Experimental Setup

We run our experiments on four different types of emotion labels, including experienced arousal, experienced valence, intended arousal and intended valence. The batch size is 256 and sequence length (T = 6, 30 seconds) for training. All trainable parameters in models are learned using the cross entropy loss on emotion class and updated by an Adam optimizer with learning rate in ranges [1e-4, 1e-3]. Early stopping is applied. We do leave-one-movie-out cross validation, i.e., 12 folds and leave extra ten percent data in each testing set as validation set. The metric we use to evaluate model performance is unweighted average recall (UAR).

3.2. Comparison Models

- **LSTM** [17] and **A**ttend**A**ffect**N**et (**AAN**) [15] is used to track the emotion trace in COGNIMUSE. We refer to these two models as our baseline. We re-implement the LSTM model ourselves while AAN is re-implemented using the author's source code.
- **Trans**f**or**m**e**r (**TFM**): Transformer has been shown to achieve state-of-the-art performance on the audio emotion recognition task [18]. We used the encoder part of the transformer for emotion classification tasks. The encoder contains 2 layers and 2 heads. We use greedy search for hidden dimensions for both linear fully connected layers and transformer in sets {32, 64}. The last hidden layer passes to the fully connected layer for emotion prediction. In our proposed end-to-end model, transformer is jointly trained as the back-end emotion classifier with different masking methods.
- **S**elf-**A**djusted mask (**SA**): We mask acoustic descriptors using self-adjusted masks (M^{bb}) as described in section 2.
- **ASM** and **ASM-SA**: We mask acoustic descriptors with linear combination of the original ASM. ASM-SA is the trainable α summation of ASM and SA masking.
- **A**djusted **A**uditory **S**aliency mask (**AAS**) and **ASMT**: Masking acoustic descriptors by AAS mask (M^{ab}), and ASMT is our proposed model.

Table 2. The unweighted average recall (UAR) and the recall in each emotion class for all models.

	Experienced Arousal								Experienced Valence							
	LSTM	AAN	TFM	ASM	AAS	SA	ASM-SA	ASMT	LSTM	AAN	TFM	ASM	AAS	SA	ASM-SA	ASMT
Low	78.57	90.44	83.98	87.31	87.35	84.91	81.28	90.95	3.47	0.79	10.09	5.52	7.99	5.36	8.68	9.31
Mid	28.82	29.27	29.27	28.89	39.95	34.54	30.93	21.60	71.44	89.15	75.78	73.95	79.87	76.27	78.19	78.98
High	0.00	0.00	6.25	5.00	6.25	3.75	0.00	26.25	49.89	41.79	52.59	58.29	54.31	53.34	56.41	58.81
UAR	38.66	40.10	42.02	40.40	44.52	41.07	37.40	**46.26**	45.50	43.90	46.15	45.92	47.20	44.99	47.76	**49.03**
	Intended Arousal								Intended Valence							
	LSTM	AAN	TFM	ASM	AAS	SA	ASM-SA	ASMT	LSTM	AAN	TFM	ASM	AAS	SA	ASM-SA	ASMT
Low	41.77	61.51	58.03	61.31	62.92	57.83	62.45	65.06	42.74	40.73	46.37	42.27	45.29	45.13	50.15	49.38
Mid	62.85	51.86	55.80	54.36	50.36	63.74	52.25	55.64	48.54	62.31	52.28	58.88	54.82	54.31	56.15	54.57
High	20.46	23.75	38.91	35.48	38.05	36.05	32.47	39.77	36.30	45.11	54.54	46.89	49.02	58.63	49.20	56.58
UAR	45.00	45.90	50.92	50.38	50.44	52.54	49.06	**53.49**	45.73	49.60	51.06	49.35	49.71	52.69	51.84	**53.51**

Table 3. The most attended feature categories.

Mask	Type	Dim	LD	MFCC	LMFB	F0	LSPF	VFU	SM	JT
AAS	Exp	All	5	2	1	4	2	3	**11***	4
SA	Intd	All	**9***	0	0	7	1	6	8	1
ASMT	Exp	Act	1	1	1	2	1	4*	2	4*
	Exp	Val	1	0	0	3	3	5*	3	1
	Intd	Act	2	1	2	0	0	3	4*	4*
	Intd	Val	1	0	1	1	2	4*	4*	3

3.3. Result and Analysis

3.3.1. Performance Comparison

Table 2 shows complete results of audio emotion classification of movies. Our proposed model ASMT obtains the best UAR in all four different types of emotion classification tasks, the improvement reaches 1.74%, 1.27%, 0.95%, 0.82% when comparing to the best of the other models on experienced arousal, experienced valence, intended arousal, intended valence, respectively. In the second column result, it shows that the use of adjusted auditory saliency masking (adjusted mask) results in a more informative input than non-adjusted case (original ASM), which is an improvement of 4.12%, 1.28%, 0.06% and 0.36% in four different emotion types respectively. This shows the importance of the use of trainable pooling mechanism to learn the mask by integrating feature sets of different types and dimensions in nature.

Furthermore, we observe that AAS performs better on the experienced emotion while SA is better on the intended emotion. Emobase2010 feature sets are extracted directly from the audio track of the movies, intuitively, modeling these content-based audio features would work best for the intended emotion recognition. However, it is quite intriguing to see that by injecting auditory saliency (AAS), which is a perceptual aspect of integrating mechanism of sensory processing [19], it would help improve the experienced emotion recognition which is closer to the audience's true emotional responses.

3.3.2. Mask Analysis

We provide an analysis on which acoustic feature categories are attended to the most in our masked representations, and these analyses are carried out in the following settings: (1) comparison between self-adjusted and adjusted auditory saliency mask and (2) masks in our proposed model. There are eight major categories in embase2010: loudness (LD), MFCC, logMelFreqBand (LMFB), F0, lspFreq (LSPF), voicingFinalUnclipped (VFU), shimmer (SM) and jitter (JT). To identify which category dominate, we retrieve the feature categories corresponding to the top largest 25% absolute value in the learned masks M^{**} for each examined model.

(1) **SA and AAS mask comparison**: First, we examine the mask within the model that improved the most for specific recognition tasks, i.e., SA on experienced and AAS on intended. The upper part of table 3 shows the counting of selected categories for the two different masks in eight feature categories. We observe the mask in SA focuses on the loudness, F0 and shimmer, while the mask in AAS enhances shimmer the most. This seems to corroborate with past findings indicating that loudness level is highly correlated to the expression of emotion [20], and the shimmer is also another important factor in the perceptual process of emotion for cross-linguistic listeners [21].

(2) **Mask in ASMT**: We then examine the masks learned in our proposed ASMT. In the lower part of Table 3, we observe that the voicingFinalUnclipped, shimmer and jitter are enhanced more than other feature categories. These three categories are closely related to voice quality. It is interesting to see that our front-end masking mechanism employed in ASMT learns to emphasize largely on the dimension of voice quality. Understanding how voice quality in audio tracks encodes emotion elicitation or influences audience affective viewing experiences will be an interesting direction to pursue.

4. CONCLUSIONS AND FUTURE WORK

In this work, we propose an end-to-end saliency-based masking transformer for emotion recognition in audio tracks. We demonstrate that by integrating auditory saliency and considering both sensory-level saliency and task-level attention, it learns a discriminative embedding that achieves the SOTA classification performances. We also observe interestingly that AAS masks are better for experienced emotion, while SA masks provide more improvement in intended emotion. In future work, we would continue to explore the connection between auditory saliency and other related tasks, such as speech enhancement or stuttering event detection, and further to extend to audio-visual multimodal framework.

5. REFERENCES

[1] D. N. Amali, A. R. Barakbah, A. R. A. Besari, and D. Agata, "Semantic video recommendation system based on video viewers impression from emotion detection," in *2018 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC)*. IEEE, 2018, pp. 176–183.

[2] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, "Corpus development for affective video indexing," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 1075–1089, 2014.

[3] M. Xu, J. Wang, X. He, J. S. Jin, S. Luo, and H. Lu, "A three-level framework for affective content analysis and its case studies," *Multimedia tools and applications*, vol. 70, no. 2, pp. 757–779, 2014.

[4] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[5] E. I. Knudsen, "Fundamental components of attention," *Annu. Rev. Neurosci.*, vol. 30, pp. 57–78, 2007.

[6] T. Parr and K. J. Friston, "Attention or salience?" *Current opinion in psychology*, vol. 29, pp. 1–5, 2019.

[7] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

[8] C.-T. Do and Y. Stylianou, "Weighting time-frequency representation of speech using auditory saliency for automatic speech recognition." in *INTERSPEECH*, 2018, pp. 1591–1595.

[9] A. Gallardo Antolín and J. M. Montero Martínez, "A saliency-based attention lstm model for cognitive load classification from speech," 2019.

[10] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 2741–2745.

[11] A. Zlatintsi, P. Koutras, G. Evangelopoulos, N. Malandrakis, N. Efthymiou, K. Pastra, A. Potamianos, and P. 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.

[12] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. 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.

[13] A. Goyal, N. Kumar, T. Guha, and S. S. Narayanan, "A multimodal mixture-of-experts model for dynamic emotion prediction in movies," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2822–2826.

[14] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[15] H. T. Phuong Thao, T. Balamurali B, D. Herremans, and G. Roig, "Attendaffectnet: Self-attention based networks for predicting affective responses from movies," *arXiv e-prints*, pp. arXiv–2010, 2020.

[16] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: an auditory saliency map," *Current biology*, vol. 15, no. 21, pp. 1943–1947, 2005.

[17] S. Sivaprasad, T. Joshi, R. Agrawal, and N. 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.

[18] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition." in *Interspeech*, 2019, pp. 2578–2582.

[19] V. Duangudom and D. V. Anderson, "Using auditory saliency to understand complex auditory scenes," in *2007 15th European Signal Processing Conference*. IEEE, 2007, pp. 1206–1210.

[20] K. R. Scherer, J. Sundberg, B. Fantini, S. Trznadel, and F. Eyben, "The expression of emotion in the singing voice: Acoustic patterns in vocal performance," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 1805–1815, 2017.

[21] D. Erickson, A. Rilliard, T. Shochi, J. Han, H. Kawahara, and K. Sakakibara, "A cross-linguistic comparison of perception to formant frequency cues in emotional speech," *COCOSDA, Kyoto, Japan*, pp. 163–167, 2008.