

Training Strategies to Handle Missing Modalities for Audio-Visual Expression Recognition

Srinivas Parthasarathy
Amazon
Sunnyvale, California
parsrini@amazon.com

Shiva Sundaram
Amazon
Sunnyvale, California
sssundar@lab126.com

ABSTRACT

Automatic audio-visual expression recognition can play an important role in communication services such as tele-health, VOIP calls and human-machine interaction. Accuracy of audio-visual expression recognition could benefit from the interplay between the two modalities. However, most audio-visual expression recognition systems, trained in ideal conditions, fail to generalize in real world scenarios where either the audio or visual modality could be missing due to a number of reasons such as limited bandwidth, interactors' orientation, caller initiated muting. This paper studies the performance of a state-of-the-art transformer when one of the modalities is missing. We conduct ablation studies to evaluate the model in the absence of either modality. Further, we propose a strategy to randomly ablate visual inputs during training at the clip or frame level to mimic real world scenarios. Results conducted on in-the-wild data, indicate significant generalization in proposed models trained on missing cues, with gains up to 17% for frame level ablations, showing that these training strategies cope better with the loss of input modalities.

CCS CONCEPTS

• **Computing methodologies** → *Neural networks*; **Supervised learning by regression**; • **Human-centered computing** → **HCI theory, concepts and models**.

KEYWORDS

expression recognition, multimodal transformers, ablation training

ACM Reference Format:

Srinivas Parthasarathy and Shiva Sundaram. 2018. Training Strategies to Handle Missing Modalities for Audio-Visual Expression Recognition. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/1122445.1122456>

1 INTRODUCTION

Expressions are a fundamental part of human communication. Humans inherently process and react to situations depending on how

well they detect others expressions. Automatic recognition of expression therefore plays an important role in building intuitive interfaces for human-machine interaction. Expression recognition systems have varied uses in fields such as health [6], security [5], customer interaction [9]. In particular, expression recognition has interesting uses in communication services where systems can detect the onset of boredom [11], extract salient regions (hotspots) in an interaction [19] or indexing [21]

Recently, the proliferation of audio and visual sensors has seen a rise in development of expression recognition systems that are trained on data from the wild. Cai et al. provide strategies to fuse audio and visual features at the feature and model level [3]. They use an ensemble architecture to extract visual cues from the face and combine with audio. For feature level fusion they use a simple concatenation of features and for model level fusion they employ a Bayesian network, showing competitive results for the task. Kollias et al. [13] provide a late fusion approach to concatenate audio and visual cues extracted through a deep convolutional neural network. Vielzof et al. study the temporal fusion of audio and visual cues using 2D and 3D CNNs. [24]. Avots et al. study cross-corpus evaluation of audio-visual expression recognition systems, combining three in the wild datasets for their training [2]. Other works have proposed ensemble approaches for expression recognition in the wild [8, 10]. These studies have established that the audio and visual cues provide complimentary information for expression detection. Furthermore, Albanie et al. successfully transferred emotion information from a teacher model trained on visual cues to a student using audio cues [1]. While these systems, attempt to bridge the gap between laboratory and realistic settings, another particular challenge arises when either input modality is completely or partially missing. Complete ablation of a modality could stem from user-initiated loss of modality, e.g. turning off the camera and partial ablations could stem from connectivity issues where frames or sequence of frames are dropped from a session. A significant question that needs further research is model robustness in the absence of input modality.

Building upon these works, this study analyzes the performance of audio-visual expression recognition systems when the input modalities are missing. While previous works have studied the effect of input modalities for other audio-visual tasks, for e.g. lip-reading [4], action recognition [14], this study understands their dynamics for expression recognition. We conduct evaluations on the Aff-Wild2 database [12]. We first establish a baseline state-of-the-art transformer architecture with a cross-modal attention layer [23]. This self-attention model intuitively learns to focus on important parts of the input signal. We illustrate the degradation in performance of this system when modalities are ablated, showing

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Woodstock '18, June 03–05, 2018, Woodstock, NY

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/10.1145/1122445.1122456>

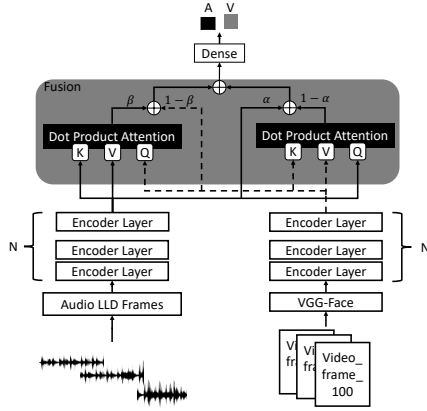


Figure 1: Baseline transformer architecture with a cross-modal attention layer for audio-visual expression detection

its susceptibility to the loss of inputs. To combat this, we also propose simple techniques to ablate data during training. We a) Ablate data at the clip level where the data is zeroed for the entire clip, b) Ablate data at the frame level where frames are either replaced by zeros or repeated. The training and evaluation strategies represent real communications scenarios that span entire loss of a modality or intermittent losses. We provide comprehensive analysis of all models evaluating them at various ablation operating points. Our results show that proposed models significantly outperform the baseline model under ablated test conditions both at the frame and clip level, while achieving near to state-of-the-art performance when tested under ideal conditions, indicating a superior generalization of the proposed models.

This paper mainly contributes by evaluating audio-visual expression recognition systems under a wider range of mismatched conditions which are present in real world scenarios.

2 PROPOSED MODEL

This work focuses on the time-continuous recognition of expressions. Given a sequence of audio-visual frames, the task is to predict the arousal (level of excitement) and valence (level of positivism) values for each frame in the sequence.

2.1 Crossmodal Transformer Architecture

This study uses a transformer architecture [23] as a baseline for audio-visual expression recognition (Figure 1). Few previous studies have explored the use of transformers for expression recognition [15, 20]. First, the audio and video inputs are independently passed through several encoder layers in two separate branches, one for the audio and one for the video. The encoder layers consist of dot-product attention layers, attending to significant regions within each modality. The dot-product attention layers work by constructing *Key* (K), *Value* (V) and *Query*(Q) triplet. The output is then calculated as

$$Y = \sigma(KQ^T)V \quad (1)$$

where σ denotes the softmax operation.

We fuse the audio and video cues using a cross-modal attention layer. This layer contains two dot-product attention components that attend to the other modality. Similar to the encoder layers, the output is obtained using Equation 1, with the main difference being

K and V are obtained from one modality while Q is obtained from the other modality. The cross-modal attention outputs are added back to the corresponding encoder representations using learnable scalar weights α and β . The overall representation from the cross-modal layer is the sum of the individual components. A dense layer is then used to get output predictions. This cross-modal attention architecture has shown competitive performance for automatic speech recognition tasks [17].

2.2 Training with missing data

To better cope with the loss of input modality, this study proposes two strategies to train with missing data (Figure 2). Our results with the baseline model (Section 4) indicate that the performance is more severely affected by the loss of visual cues than and the absence of audio cues negligibly affects the performance. Therefore the focus of the paper is to train by ablating the visual modality. In the first method we ablate data at the clip level - **Clip-Zero**. Each clip is zeroed out with a given probability p_{clip} . This corresponds to zeroing out an entire sequence of frames. In the second method we ablate data at the frame level. At the frame level, we further subdivide the training into two categories. Each frame in a sequence is either zeroed out **Frame-Zero** or repeated **Frame-Repeat** with a probability p_{frame} . Note, that these training techniques can be viewed as special forms of data augmentation. Frame-zero and Clip-zero can be viewed as extreme cases of input feature dropout where the entire input at the frame or sequence level is dropped out. With these viewpoints, we expect the model to better generalize to different test conditions. Overall these training strategies imitate the different scenarios in which data might be missing during evaluation and real use cases.

3 EXPERIMENTAL SETUP

3.1 Dataset

This study uses the the Aff-Wild2 database [12]. The dataset consists 558 videos sourced from YouTube, mostly containing single person expressions. Videos are captured in the wild conditions with hugely varying acoustic and visual environments, making it ideal for this study. All the videos are captured at a frame rate of 30 frames per second (fps) with around 2.8 million video frames in the entire dataset. The dataset is annotated for 3 main behavior tasks (basic expression classification, valence-arousal estimation and action unit detection.) This paper focuses on frame level valence-arousal estimation. Each frame in the dataset is annotated by 4 or 8 annotators for valence and arousal score in a continuous scale ranging between [-1, 1]. Further, the dataset is divided into speaker independent train, validation and test subsets consisting of 350, 70 and 138 videos respectively.

3.2 Features and Preprocessing

3.2.1 Video. Following previous studies, we extract bounding boxes for to detect frontal faces every video frame [13]. The face detector is based on an SSD face detector with a ResNet architecture [16]. The face bounding boxes are further cropped and resized to dimensions of 96x96x3. Pixel intensities are normalized between [-1, 1]. To extract features from the face frames, we use a pretrained VGG-Face model [18]. Representations from the first fully connected layer

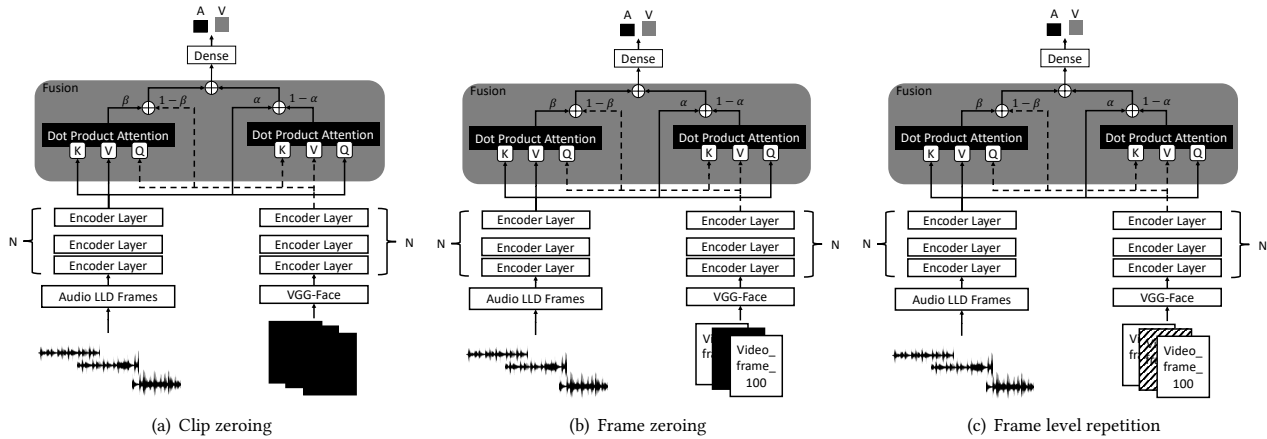


Figure 2: Proposed ablation strategies. All ablations are conducted on the visual modality

with a dimension of 4096 is fed as input to the video branch of the transformer.

3.2.2 *Audio*. For the audio features we use the features introduced at the para-linguistic challenge in Interspeech 2013 [22]. These features have shown superior performance for audio expression recognition tasks. This feature set consists of a variety of acoustic features grouped into spectral, energy related and voice features such as *Mel Frequency Cepstral Coefficients* (MFCC), loudness and fundamental frequency (F0). Features are extracted over 30ms windows with a 10ms step. The feature set consists of 65 *low-level descriptors* (LLD) and Z-normalized over the train set. To synchronize the audio and video features the audio frames are further downsampled by a factor of 3.3 and concatenated over a 2s window to increase context and match the dimension of the video branch. Overall, the audio branch of the transformer receives a 3900-dimensional vector.

3.3 Experimental framework

Arousal and valence predictions are made on sequences of 100 frames which corresponds to 3s of the input video. Audio and video encoders branches of the transformer contain 2 self-attention layers with 512 nodes and 4 heads for multihead attention. For the proposed models, the probability of ablation at both clip level p_{clip} and frame-level p_{frame} is fixed at 0.5. Since the test data set is not released publicly, all models are optimized on a part of the train set and best results are evaluated on the validation set. We use an Adam optimizer with an initial learning rate of $1e-4$. Since this is a sequence task, models are trained and evaluated using the *Concordance Correlation Coefficient* (CCC) metric.

4 RESULTS

To evaluate the proposed models, we propose various test conditions. Similar to the train conditions, we ablate the evaluation data inputs at the clip level - **Clip-Zero** and frame level - **Frame-Zero** and **Frame-Repeat**, at various probabilities of missing data (Figure 3 and Figure 4). For Clip-Zero we evaluate using probabilities $p \in 1.0, 0.7, 0.5, 0.3, 0.0$ where the extremes correspond to the complete absence or complete presence of the modality. For

Frame-Zero and Frame-Repeat we use more severe probabilities $p \in 1.0, 0.95, 0.90, 0.85, 0.0$.

We first observe that under ideal conditions ($p = 0.0$) the baseline model achieves CCC values of 0.393 and 0.522 for valence and arousal respectively. Under these conditions all proposed models are able to match baseline performance in most cases. Note that predicting valence suffers more than predicting arousal, indicating a greater dependence of valence on the visual modality.

4.1 Audio Ablations

For audio ablations, all models in Figure 3 show little variation in performance at various probabilities and various test conditions ($< 1\%$ absolute drop in performance). This indicates that the proposed models are more dependent on the rich video modality for the Aff-Wild2 database. While this may be particular to the Aff-Wild2, as shown by other studies, the ablation training methodology holds [25]. The proposed models cope with baseline performance even with half the video data removed. Note that a model trained only on the video stream, without cross-modal attention achieves CCC values of 0.381 for valence and 0.489 for arousal.

4.2 Video Ablations

Perhaps the more interesting results are with evaluations on ablating the visual modality (Figure 4). Evaluations with Clip-Zero (Figure 4(a)) indicate that the model trained with Clip-Zero, overall handles ablations better than the baseline model. The performance is significantly better at high probabilities especially for arousal. Note that at $p = 1.0$, the baseline model's performance is decreased by almost 100%. In comparison the Clip-Zero model performs similar to a model trained only on the audio-stream without cross-modal attention (CCC scores of -0.060, 0.284 for valence and arousal respectively). This indicates that after training, the model is more dependent on audio cues, and automatically switches to inference using the available audio modality. Frame-Zero models show superior performance in certain scenarios for arousal when evaluated with clip-level ablations.

Evaluations with Frame-Zero (Figure 4(b)), Frame-Repeat (Figure 4(c)) ablations indicate the necessity for training the model with matching conditions. Even at $p = 0.95$, where a large number

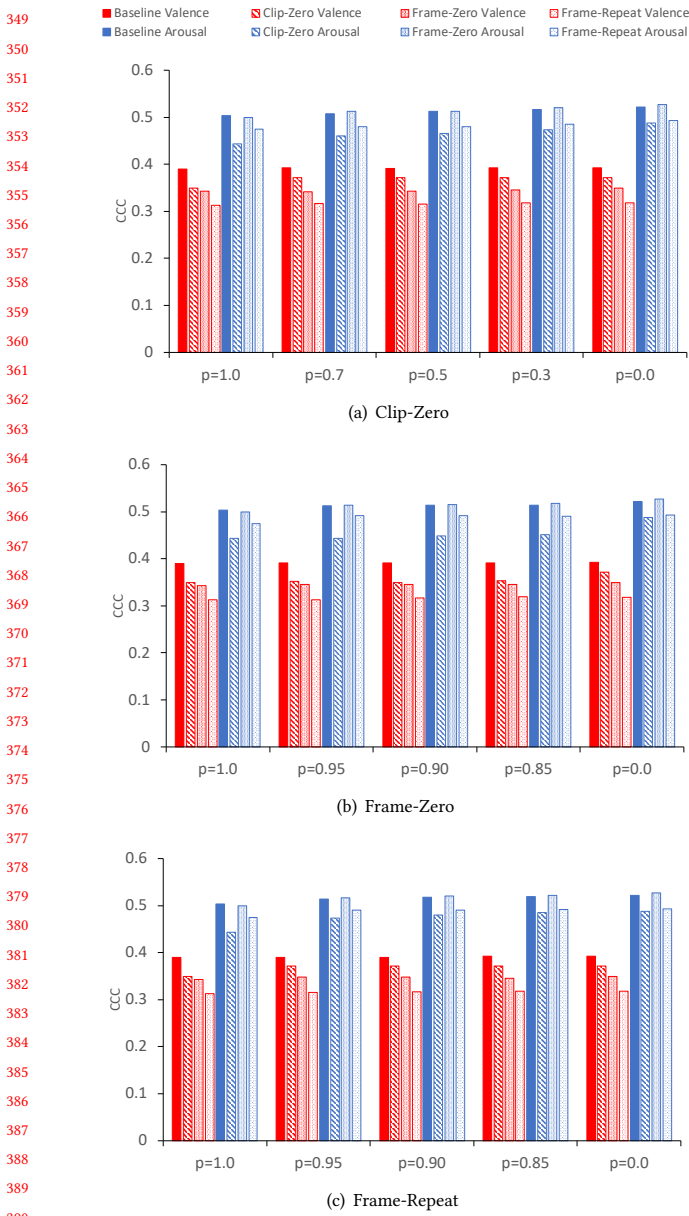


Figure 3: Results of training with missing data, evaluated with ablations of the audio modality at various probabilities for Clip-Zero, Frame-Zero and Frame-Repeat

of frames is missing the baseline model is able to recover significant performance (CCC-V=0.113, CCC-A=0.235 for Frame-Zero and CCC-V=0.285, CCC-A=0.410 for Frame-Repeat). But the proposed Frame-Zero and Frame-Repeat models do even better, and significantly outperform the baseline models for both attributes. Proposed models achieve gains of up to 17% and quickly saturate to performance when evaluated at ideal conditions $p = 0.0$. These results indicate the model is able to learn the expression from a few video frames in the sequence. The performance for Frame-Repeat evaluations is less pronounced than Frame-Zero evaluations but

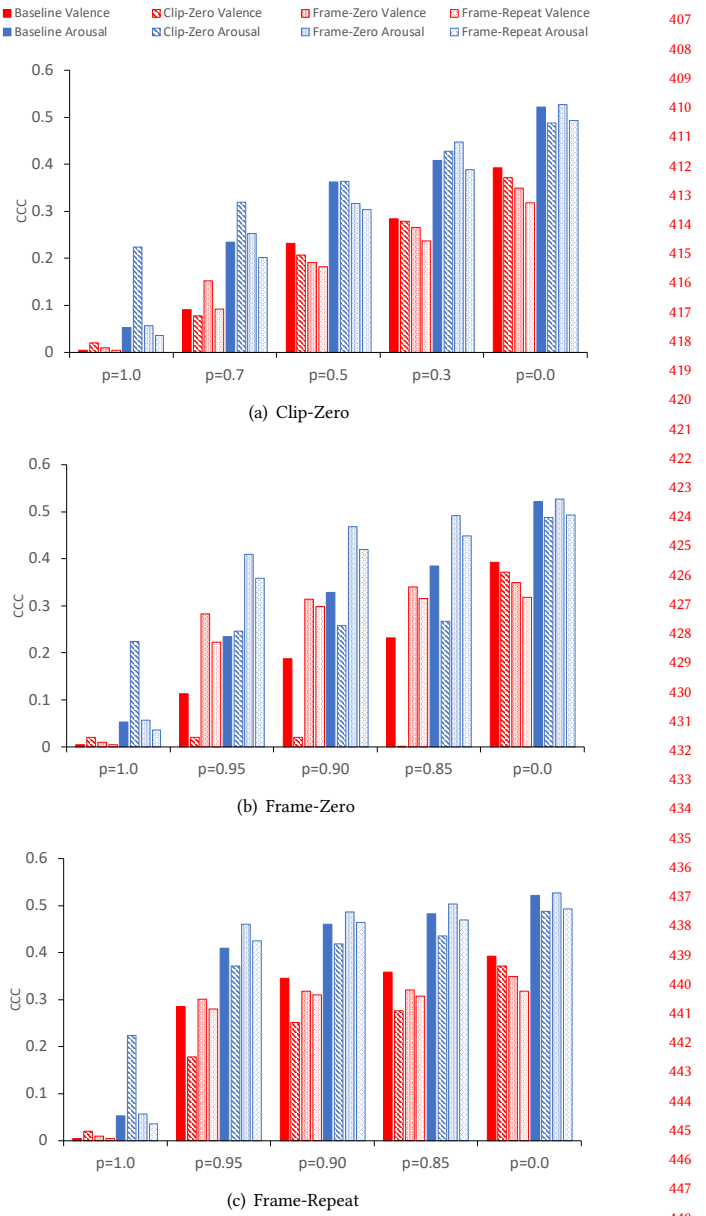


Figure 4: Results of training with missing data, evaluated with ablations of the video modality at various probabilities for Clip-Zero, Frame-Zero and Frame-Repeat

show similar trends. Overall these results show the generalization of the proposed training techniques across various test conditions.

5 CONCLUSION

This study analyses the performance of a state-of-the-art transformer model, with attention, for audio-visual expression recognition in the wild. We show that the model suffers when input data is ablated. We propose a training strategy to handle missing input modalities. We randomly ablate visual cues by zeroing them out at the clip-level or zeroing or repeating them at the frame level. Other

works have proposed ensemble approaches for expression recognition in the wild [7, 10]. Our work can be viewed as a particular form of ensembling geared towards real world. Results indicate proposed models perform better than the baseline model when data is missing during evaluation without significant loss in performance under ideal conditions. These results indicate the superior generalization of the proposed models for various test cases addressing challenges for a real use case.

Our future work will explore further ablation strategies for training. In particular we will explore subclip level ablations where portions of the clip will be zeroed. Further different probability distributions can be utilized to ablate data at the frame level. We will also explore strategies where only the faces are ablated and the background scene information can be utilized rather than zeroing.

REFERENCES

- [1] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. 2018. Emotion recognition in speech using cross-modal transfer in the wild. In *Proceedings of the 26th ACM international conference on Multimedia*. 292–301.
- [2] Egils Avots, Tomasz Sapiński, Maie Bachmann, and Dorota Kamińska. 2019. Audiovisual emotion recognition in wild. *Machine Vision and Applications* 30, 5 (2019), 975–985.
- [3] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O’Reilly, Shizhong Han, Ping Liu, Min Chen, and Yan Tong. 2019. Feature-level and model-level audiovisual fusion for emotion recognition in the wild. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 443–448.
- [4] J.S. Chung, A. Senior, O. Vinyals, and A. Zisserman. 2017. Lip reading sentences in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*. Honolulu, HI, USA, 3444–3453. <https://doi.org/10.1109/CVPR.2017.367>
- [5] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette. 2008. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication* 50, 6 (June 2008), 487–503.
- [6] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T.F. Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication* 71 (July 2015), 10–49. <https://doi.org/10.1016/j.specom.2015.03.004>
- [7] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. Mower Provost. 2017. Progressive Neural Networks for Transfer Learning in Emotion Recognition. In *Interspeech 2017*. Stockholm, Sweden, 1098–1102. <https://doi.org/10.21437/Interspeech.2017-1637>
- [8] John Gideon, Biqiao Zhang, Zakaria Aldeneh, Yelin Kim, Soheil Khorram, Duc Le, and Emily Mower Provost. 2016. Wild wild emotion: a multimodal and ensemble approach. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 501–505.
- [9] Purnima Gupta and Nitendra Rajput. 2007. Two-stream emotion recognition for call center monitoring. In *Eighth Annual Conference of the International Speech Communication Association*.
- [10] Ping Hu, Dongqi Cai, Shandong Wang, Anbang Yao, and Yurong Chen. 2017. Learning supervised scoring ensemble for emotion recognition in the wild. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 553–560.
- [11] Eun-Hye Jang, Byoung-Jun Park, Mi-Sook Park, Sang-Hyeob Kim, and Jin-Hun Sohn. 2015. Analysis of physiological signals for recognition of boredom, pain, and surprise emotions. *Journal of physiological anthropology* 34, 1 (2015), 25.
- [12] Dimitrios Kollias and Stefanos Zafeiriou. 2018. Aff-Wild2: Extending the Aff-Wild Database for Affect Recognition. *arXiv preprint arXiv:1811.07770* (2018).
- [13] Dimitrios Kollias and Stefanos Zafeiriou. 2019. Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace. *arXiv preprint arXiv:1910.04855* (2019).
- [14] H. Lee, C. Lin, P. Hsu, and W. H. Hsu. 2019. Audio Feature Generation for Missing Modality Problem in Video Action Recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3956–3960.
- [15] Yuanchao Li, Tianyu Zhao, and Tatsuya Kawahara. 2019. Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning.
- [16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. *Lecture Notes in Computer Science* (2016), 21–37. https://doi.org/10.1007/978-3-319-46448-0_2
- [17] Georgios Paraskevopoulos, Srinivas Parthasarathy, Aparna Khare, and Shiva Sundaram. 2020. Multimodal and Multiresolution Speech Recognition with Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2381–2387. <https://doi.org/10.18653/v1/2020.acl-main.216>
- [18] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep Face Recognition. In *British Machine Vision Conference*.
- [19] S. Parthasarathy and C. Busso. 2019. Predicting Emotionally Salient Regions using Qualitative Agreement of Deep Neural Network Regressors. *IEEE Transactions on Affective Computing* To appear (2019). <https://doi.org/10.1109/TAFFC.2018.2878715>
- [20] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating Multimodal Information in Large Pretrained Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2359–2369. <https://doi.org/10.18653/v1/2020.acl-main.214>
- [21] Gaël Richard, Shiva Sundaram, and Shrikanth Narayanan. 2013. An overview on perceptually motivated audio indexing and classification. *Proc. IEEE* 101, 9 (2013), 1939–1954.
- [22] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim. 2013. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In *Interspeech 2013*. Lyon, France, 148–152.
- [23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
- [24] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. 2017. Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 569–576.
- [25] Yuan-Hang Zhang, Rulin Huang, Jiabei Zeng, Shiguang Shan, and Xilin Chen. 2020. M^3T : Multi-Modal Continuous Valence-Arousal Estimation in the Wild. [arXiv:cs.CV/2002.02957](https://arxiv.org/abs/2002.02957)