Deep Learning for Audio Visual Emotion Recognition

T. Hussain[1], W. Wang[2], N. Bouaynaya[3], H. Fathallah-Shaykh[4], and L. Mihaylova[1]
[1]Department of Automatic Control and Systems Engineering, The University of Sheffield, UK
[2]Centre for Vision Speech and Signal Processing, The University of Surrey, UK
[3]Department of Electrical and Computer Engineering, Rowan University, USA
[4]School of Medicine, University of Alabama at Birmingham, USA
Email: {tassadaq.hussain@gmail.com, w.wang@surrey.ac.uk,
bouaynaya@rowan.edu, hfshaykh@uabmc.edu, l.s.mihaylova@sheffield.ac.uk}

***Abstract*—Human emotions can be presented in data with multiple modalities, e.g. video, audio and text. An automated system for emotion recognition needs to consider a number of challenging issues, including feature extraction, and dealing with variations and noise in data. Deep learning have been extensively used recently, offering excellent performance in emotion recognition. This work presents a new method based on audio and visual modalities, where visual cues facilitate the detection of the speech or non-speech frames and the emotional state of the speaker. Different from previous works, we propose the use of novel speech features, e.g. the Wavegram, which is extracted with a one-dimensional Convolutional Neural Network (CNN) learned directly from time-domain waveforms, and Wavegram-Logmel features which combines the Wavegram with the log mel spectrogram. The system is then trained in an end-to-end fashion on the SAVEE database by also taking advantage of the correlations among each of the streams. It is shown that the proposed approach outperforms the traditional and state-of-the art deep learning based approaches, built separately on auditory and visual handcrafted features for the prediction of spontaneous and natural emotions.**

Keywords: Deep learning, convolutional neural networks, emotion recognition, audio and visual data.

I. Introduction

Emotion recognition is the process of identifying human emotions, typically from facial expressions, via computational methods. Unlike humans who are capable in recognizing emotions, computational approaches are limited in recognizing low-level and high-level feature from different modalities, such as audio, images, and videos, in a human-like manner. The research area of human emotion recognition is increasingly active over the past few years. However, recognizing emotions in real acoustic environments and under different speech production conditions remains a key challenge. Numerous unimodal and multimodal emotion recognition methods have been proposed and widely used as a pre-processor in speech-related applications, for example, human-computer interaction [1], robots [2], mobile services [3], call centers [4], computer games [5], and psychological assessment [6] [7]. A number of survey papers for human emotion recognition have been published in recent years. Among these survey papers, Ververidis and Kotropoulos [8], Ayadi et al., [9], Koolagudi and Rao [10], and Basu et al. [11] are the well-known articles that discussed the use of speech modality for human emotion recognition. Despite of these extensive works, the recognition performance of existing methods is still limited in real-world settings.

Apart from conventional approaches, data-driven deep learning (DL)-based solutions, which are gradient-based neural architectures, have proven useful in overcoming some limitations of conventional emotion recognition frameworks. With the recent development of DL algorithms in behavioral signal processing and affective computing, the DL-based emotion recognition algorithms have received significant attention. Notable DL-based emotion recognition approaches include Long-Short Time Memory architectures [12] [13] [14]), deep neural network (DNN) [15], convolutional neural network (CNN) [16] [17] [18], and bidirectional long short-term memory (BLSTM) [19]. Among the DL-based models, CNNs have been shown to be effective in detecting emotions, due to its capability in characterizing local temporal-spectral structures of speech and audio signals, as well as its generalisation ability and recognition accuracy. In recent years, multimodal DL frameworks (i.e., audio-visual (AV) and audio-text) have been utilized to learn multiple levels of representations that correspond to different levels of abstraction, where each level forms a hierarchy of emotions. Notable DL-based multimodal emotion recognition approaches include Ouyang et al [20], Zhang et al. [21], Kansizoglou et al. [22], Ma et al. [23], Wang et al. [24], and Schoneveld et al. [25].

Despite the state-of-the-art performance achieved by DL models for emotion recognition, DL models are viewed as deterministic functions [26], and as a

result are unable to understand the model behaviour, a critical part of any predictive system's output. This can have disastrous consequences, especially when the output of such models is then fed into higher-level decision-making procedures. These include emotion recognition and recommendation systems in the medical domain, autonomous vehicles, and marketing. In these scenarios, quantifying model to adapt the subsequent decision-making process might be key to preventing unintended behaviour.

The initial step in any AV emotion recognition system is to pre-process the visual data to detect the face followed by the removal of background and non-facial areas. The Viola-Jones (VJ) face detector [27] is a well-known and commonly used method for near-frontal face identification that is both robust and computationally simple. Although face identification is the only process required to enable feature learning, subsequent face alignment utilising the coordinates of localised landmarks can improve the AV emotion recognition performance significantly. Because different face detectors often output distinct face bounding boxes with varied sizes and centre shifts, there still remains uncertainty whether a facial landmark recognition algorithm could produce reliable findings without relying heavily on the face detection results. Moreover, we notice that while identifying the boundary boxes, there are still ambiguities. In this paper, we extend the AV emotion recognition research and proposed a new bounding box regression loss for learning bounding box transformation and localisation variance simultaneously for robust emotion recognition performance in unseen/real-time environments. The new learned localisation variance manages to merge neighbouring bounding boxes to further enhance localisation efficiency and improves the AV emotion recognition accuracy. In this work, we employ a deep neural-based multimodal (AV) architectures to estimate emotion states.

Our main contributions in this paper are the following:

1. An audio-visual framework for emotion recognition is proposed based on intermediate level fusion of audio and image features.

2. To improve audio representation, we used the PANNs system to generate the Wavegram and log-mel spectrogram audio features, where the PANNs [28] system was pre-trained on AudioSet [29], a large scale audio dataset containing 2,063,839 audio clips, designed originally for audio tagging tasks.

3. A hybrid feature extraction and learning method is proposed to fuse efficiently the audio and video features for the purposes of emotion recognition. The framework combines hand-crafted speech features, such as Mel-frequency cepstral coefficient (MFCC), log MFCC (logMFCC) with Wavegram and log-mel-Wavegram features for robust emotion recognition.

The remainder of the paper is divided into the sections below. Section 2 discusses related work on unimodal and audio-visual emotion recognition using deep neural networks. Section 3 presents a detailed description of the proposed DL-based AV emotion recognition framework. Section 4 presents the results and finally Section 5 concludes this work.

II. Related Work

A. Unimodal Emotion Recognition

The majority of research related to emotion recognition primarily concentrate on six facial emotions, such as, happy, sorrow, disgust, anger, fear, and surprise. The recognition of these basic facial expressions was based on Ekman's long study [30], which showed that these basic facial expressions are universally seen by humans across cultures. However, non-basic emotions account for the majority of emotion manifestations in human-to-human communication. Furthermore, the majority of existing emotion recognition systems are unimodal: the system only processes speech data or face images [31]. In recent years, multimodal affect analysis has received a lot of attention, however, a very limited research has been done to exploit the audio-visual cues for emotion recognition tasks.

B. Multimodal Emotion Recongition

Many recent multimodal emotion recognition experiments have taken advantage of a synergistic combination of different modalities. The majority of recent research has focused on fusing audio-visual (AV) data for automatic emotion recognition, such as merging voice and facial expression. For example, a combination of audio and visual features are studied by Ngiam et al. [32] using a Multimodal Deep Autoencoder (MDAE) framework. The MDAE model was fine-tuned to minimise the reconstruction error of both modalities after a bimodal deep belief networks (DBNs) was learned to initialise the deep autoencoder. Hu et al. [33] introduced the Recurrent Temporal Multimodal Restricted Boltzmann Machine (RTMRBM) as a temporal multimodal network to represent AV sequences. Gesture recognition is another job for which DNNs have been utilised. The authors of [34] detect motions using skeletal information and RGB-D pictures. DBNs are used to process skeletal features, while a 3D CNN is used to handle the RGB-D data. A Hidden Markov Model (HMM) is stacked to exploit the temporal information in the data.

A combined analysis of speech and facial expressions studied by De Silva et al. [35] and Chen et al. [36] by developing a rule-based decision level fusion technique. Further, boosting approaches were utilised

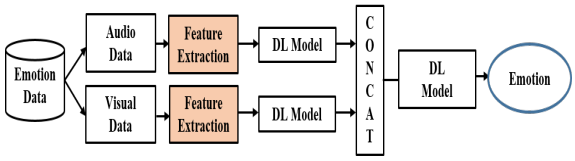

Figure 1: Block Diagram of Conventional Deep Learning-based Audio-Visual Emotion Recognition Framework

by Huang et al. [37] to compute adaptive weights for audio and visual features automatically. An emotion-specific comparison of feature-level and decision-level fusion was reported by Busso et al. [38], by utilising an AV database including four emotions, i.e. grief, rage, happiness, and neutral mood, all of which were purposely posed by an actor. They discovered that feature-level fusion was better at distinguishing between angry and neutral mood in their corpus, whereas decision-level fusion performed better for happiness and sadness. The authors in [38] came to the conclusion that the best fusion approach is determined by the application. Apart from speech and facial expression, the thermal distribution of infrared images is also integrated into a multimodal recognition system [39] by taking into account the fact that infrared images are relatively unaffected by lighting conditions, which is one of the major challenges in facial image analysis.

III. Proposed Method

The overall architecture of the proposed system is shown in Fig. 1. In this section, we first briefly discuss the audio and visual modules and how these modules are trained independently to acquire the corresponding representation for audio and visual modalities. Next, we discuss how to jointly optimise the AV framework for emotion recognition using a customized loss function.

A. Audio Network

Extracting features from data is one of the crucial steps in typical machine learning algorithms. For audio features, we extracted Wavegram and Wavegram-Logmel features as described in [28] which were initially proposed for AudioSet tagging. A Wavegram is a type of feature learned directly from speech waveforms using a one-dimensional CNN, which can be considered as a time-frequency representation of speech, with time axis representing the time frames, while frequency axis representing the frequencies derived from the channels in CNN. Sounds with various pitch shifts can belong to the same class, hence frequency patterns are significant for emotion pattern recognition. A Wavegram can help capture such frequency information, thus may outperform hand-crafted log mel spectrograms. Wavegrams may thus be used as input features, instead of log mel spectrograms, resulting in a Wavegram-CNN system. The Wavegram can also be combined with the log mel spectrogram, leading to the Wavegram-Logmel-CNN system, as illustrated in Fig. 2.

As described in [28], a one-dimensional CNN is applied to a time-domain waveform to create a Wavegram. The wavegram first reduces the size of the input waveform by applying a convolutional layer of filter length 11 and stride 5. This instantly reduces the length of the inputs by a factor of 5 to save memory. Following that, three convolutional blocks are used, each of which is comprised of two convolutional layers with dilations of 1 and 2, respectively, to enhance the receptive field of the convolutional layers. A downsampling layer with stride 4 follows each convolutional block. We downsample a 48 kHz audio waveform to 32 kHz to generate 32000/5/4/4/4= 100 frames of features per second to have a similar configuration as described in [28]. The size of the output of the one-dimensional CNN is T x C, where T is the number of frames and C is the number of channels. We transform this output into a T-dimensional tensor by dividing C channels into C/F groups, each with F frequency bins. This tensor is known as a Wavegram. By inserting F frequency bins in each of the C/F channels, the Wavegram learns frequency information. For AV emotion recognition, we combine the Wavegram and Wavegram-Logmel features of audio with the visual network as described in next section, with bounding box regression to compare the Wavegram and log mel spectrogram based systems equitably.

B. Visual Network

Besides auditory information, the proposed system also exploits visual information to get not only high-level information about speech and non-speech (i.e., silence) regions of an utterance, but also fine-grained information about mouth articulation. Although improvements were shown for all AV emotion recognition systems, emotions that are easier to distinguish visually were the ones that improved the most with an AV-SE system. Different from previous works, which utilizes traditional face recognition pipeline to detect subjects faces using bounding boxes and uses the pixel intensities from the cropped faces/bounding boxes in combination with audio to jointly train a model for emotion prediction. The current work utilizes a bounding box regression method to quantify transformation at the classification stage.

The visual network pipeline is divided into two steps. The Faster R-CNN was proposed to lower the computational cost of proposal generation. It is made up of two modules. The first module, known as the Regional Proposal Network (RPN), is a fully convolutional network that generates object proposals

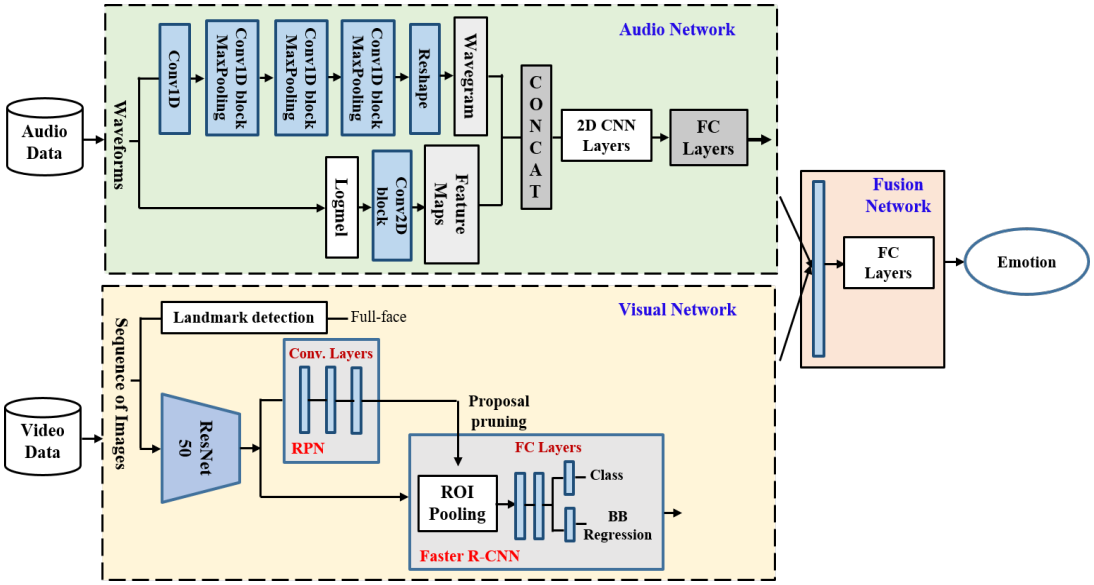

Figure 2: Block Diagram of Proposed Audio-visual (AV) Emotion Recognition Framework

that are input into the second module, i.e. the Fast R-CNN detector, whose aim is to refine the proposals. The main concept is that the RPN and Fast R-CNN detectors share the same convolutional layers up to their own fully connected layers. The image is now just processed by the CNN once to generate and then refine object recommendations. More crucially, because convolutional layers may be shared, a very deep network (e.g., ResNet [40]) can be used to create high-quality object proposals.

The convolution layers of a pre-trained network are followed by a 3 x 3 convolutional layer in the RPN. This refers to converting a large spatial window or receptive field in the input picture (e.g., 228 x 228 for VGG16) to a low-dimensional feature vector at a centre stride (e.g., 16 for VGG16). Then, for the classification and regression branches of all spatial windows, two 1 x 1 convolutional layers are added. The classifier in this case determines the likelihood of a proposal containing the target object and the regression is used for regressing the proposal coordinates.

C. Training the Proposed Audio-Visual Emotion Recognition Model

Initially, two separate networks can be trained to update the parameters of the audio and visual networks, producing more discriminative audio and visual features. We then merge the outputs of the audio and visual networks as the input to the fusion network, such as $F_i = [A_i, V_i]$. To train the fusion network, the outputs of the audio and visual networks are integrated to learn the joint representation, such as $F_i = [A_i, V_i]$, where A_i and V_i are the audio and visual features. The integrated features are subsequently forwarded to fully-connected (FC) layers to predict the emotion labels.

The audio network is optimised with a binary cross-entropy loss function $\mathcal{L}_a$, which is defined as

$$\mathcal{L}_a = -(y_i \log(p_i) + (1 - y_i) \log(1 - p_i)), \quad (1)$$

where y_i is the ground truth label of class i and p_i is the predicted probability that the emotion belongs to class i. Different from the audio network, the visual network which is based on a Faster R-CNN [41] architecture, combines the classification loss $\mathcal{L}_{cls}$ with the bounding box regression loss $\mathcal{L}_{bb}$.

The network for visual features extraction has a cost function $\mathcal{L}_v$ which combines the losses of classification $\mathcal{L}_{cls}$ and bounding box regression $\mathcal{L}_{cls}$:

$$\mathcal{L}_v = \mathcal{L}_{cls} + \mathcal{L}_{bb} \quad (2)$$

The overall loss function $\mathcal{L}_v$ can also be written as follows [41]

$$\mathcal{L}_v = \frac{1}{M_{cls}} \sum_i \mathcal{L}_{cls}(p_i, y_i) + \frac{\lambda}{M_{bb}} \sum_i y_i L_1(y_i, p_i) \quad (3)$$

over an M_{cls} number of classes in the faster R-CNN and M_{bb} classes from the regression bounding boxes, L_1 is the regression loss, with λ being a weighting coefficient allowing to give different importance to one of the two loss functions.

Optimising directly the loss shown in Eq. (3) can result in certain emotions being wrongly recognised into polarity-opposite groups. We formulate a more general loss function $Loss$ in this work. The proposed model is trained in an end-to-end manner, as shown in Fig. 1. The model is initially optimized using a single loss function, and next we add the audio and video

loss functions with a regularisation term to obtain the final results with the following loss function

During the fusion process, the intended training objective can be expressed as a sum of losses from all conceivable combinations of modality-specific models:

$$Loss = \mathcal{L}_{(a,v)} + \alpha\mathcal{L}_a + \beta\mathcal{L}_v, \quad (4)$$

where $\mathcal{L}_{(a,v)}$ is the combined loss of audio and visual framework as shown in Eq. (1) and Eq. (2), and α and β are the weights for the regularisation terms, respectively.

IV. Performance Validation and Evaluation

A. Experimental Setup

We evaluate the performance of the proposed system using the Surrey Audio-Visual Expressed Emotion (SAVEE) database [42], which was created for an automated emotion recognition system. The dataset contains 480 British English utterances recorded by four male actors in seven distinct emotion, such as anger, disgust, fearful, happy, neutral, sad and surprised, respectively. We chose 410 utterances to train the frameworks and 10 random utterances from each emotion category to test the performance of the framework. For each emotion, the sentences were taken from the normal TIMIT corpus and phonetically balanced. The data was captured, analysed, and labelled with high-quality audio-visual equipment where the speaker was facing the camera.

To analyse human facial expressions and emotions, facial landmark annotation is initially prepared to detect human faces with greater accuracy. It mainly employs facial landmark points to determine the density of an object inside a given area which aids in a better understanding of each point motion in the movement trajectory of the targeted object. Since SAVEE did not provide any information regarding face annotation and facial landmarks, we annotated human faces and bounding boxes using landmark annotation tools. Because the speakers were filmed from the front with only frontal pose, a number of APIs and annotation tools can be used to detect their faces and label bounding boxes.

For the SAVEE dataset, we label the bounding boxes for the recognizable face using Dlib [43] face detection libraries which returns the top-left corner with width and height of the region of interest (ROI). These ROI coordinates are used as the true bounding box labels to fine-tune the bounding boxes using Faster R-CNN. The annotations were prepared in a similar way to the PASCAL VOC dataset [44] where additional flags, such as blur, expression, illumination, occlusion, pose, and invalid, were set to 0. More details about flags can be found in [44].

B. Baseline Systems

To evaluate the effectiveness of the proposed model, we first compare the performance of audio- and visual-only emotion recognition systems. We then developed two simple multimodal systems for performance comparison: (i) a conventional AV emotion recognition framework without face detection trained end-to-end, and (ii) a multimodal baseline framework with face detection to only use cropped images of the speaker face as an input to the visual network alongside the audio network. In addition, two state-of-the-art end-to-end frameworks, such as DNN-based multimodal emotion recognition model [45] and the attention-based AV model [46], were trained for emotion recognition as baseline systems.

C. Performance Comparison

First, we compare the performance of unimodal systems i.e., audio- and visual-only, for emotion recognition under speaker-independent conditions. Table 1 lists the accuracy of different CNN-based audio-only and visual-only frameworks. For audio network, we adapted a pre-trained PANNs for emotion recognition. As described in [28], PANNs was originally proposed for audio-tagging and was trained using a large Audioset tagging dataset (527-classes) [29] which can easily be adapted to a variety of audio pattern recognition tasks [28]. Instead of training PANNs from scratch, we fine-tuned a PANNs framework for an emotion recognition task, where all PANNs parameters, except the final FC layer, are initialised from the PANNs, and we fine-tuned the framework by adding three additional FC layers for an emotion recognition task (7-classes).

The PANNs is used to calculate the embedding features of audio waveforms. The embedding features are then fed into a classifier, such as a fully-connected neural network, as input. The settings of the PANNs are frozen and not trained when the emotion recognition system is trained. Only the parameters of the embedding features-based classifier are learned. The PANNs is used to extract features. The extracted embedding features are used to create a classifier. For visual network, we utilised the MMdetection toolbox [47] training pipeline to train and evaluate the Faster R-CNN for face detection. As a backbone, a ResNet-50 [40] was employed without the last FC layer.

Table I first presents the performance comparison of audio- and visual-only systems for emotion recognition. For these systems, CNN frameworks are trained using distinct audio and visual features. For example, $\text{CNN}_{\text{Wavegram-Logmel}}$ and $\text{CNN}_{\text{Wavegram}}$ were trained using combined Wavegram and Wavegram-Logmel, and Wavegram features, respectively. Similarly, CNN_{fd} for visual network was trained by only considering facial landmark region. From Table I,

Table I: PERFORMANCE COMPARISON OF AUDIO- AND VISUAL-ONLY SYSTEMS.

Modality	Framework	Accuracy
Audio	CNN	59.3
	$CNN_{wavegram}$	62.2
	$CNN_{Wavegram\text{-}Logmel}$	65.5
Visual	CNN	39.1
	CNN_{fd}	31.9
	Faster R-CNN	33.8

Table II: PERFORMANCE COMPARISON OF THE PROPOSED AV SYSTEM WITH BASELINE SYSTEMS.

Modality	Framework	Accuracy
Baseline (A + V)	DNN	49.8
	VAANet	63.8
A + V	CNN	59.3
	CNN_{fd}	61.4
	Faster R-CNN_{bb}(Ours)	63.1

we can see that audio-only CNN framework considering Wavegram and Wavegram-Logmel features performed exceptionally well. On the other hand, CNN framework trained for visual-only modality considering the whole image performed well compared to the CNN's with face detection (CNN_{fd}) and Faster R-CNN. Table II shows the comparison of our proposed AV system with baseline systems. The results are highly dependent on the extracted audio and visual features and also on the testing data sets.

The baseline systems were originally proposed for regression task where the goal was to determine arousal and valence. In this work, we modified the baseline systems to accommodate them for emotion classification task utilising the same configuration reported in [46] and [45]. It is noted that the audio network in Table II was trained using combined Wavegram and Wavegram-Logmel features which proved effective for emotion recognition. When comparing overall performance, we can see that the proposed AV emotion recognition system with bounding box regression achieved comparable performance to one of the baseline systems (VAANet [48]) and outperformed other baseline systems and CNN-based AV emotion recognition systems with reasonable margin.

V. CONCLUSIONS

We have presented an end-to-end emotion recognition approach based on audio and visual data. Different from conventional audio and visual emotion recognition framework, we used Wavegram and Wavegram-LogMel features to train an audio network, and a Faster R-CNN for visual network in an end-to-end manner with a customised loss function. The Faster R-CNN not only improved the localisation performance while estimating the bounding boxes during face detection, but also increased the overall accuracy of the AV emotion recognition system. The proposed framework was trained using the SAVEE dataset with a modified cross-entropy loss, and the results show that the proposed system outperforms state-of-the art AV emotion recognition approaches.

In the future, we will use additional loss functions to improve the performance of the proposed AV emotion recognition for larger datasets and speaker independent conditions, and evaluate the proposed system under different uncertainties (noises, adversarial attacks both in the audio and in the video data).

Acknowledgements. We are grateful to EPSRC for funding this work through the EP/T013265/1 project NSF-EPSRC: "ShiRAS. Towards Safe and Reliable Autonomy in Sensor Driven" and the support for ShiRAS by the National Science Foundation under Grant USA NSF ECCS 1903466. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[2] X. Huahu, G. Jue, and Y. Jian, "Application of speech emotion recognition in intelligent household robot," in *Proc. of the International Conf. on AI and Computational Intelligence*, vol. 1. IEEE, 2010, pp. 537–541.

[3] W.-J. Yoon, Y.-H. Cho, and K.-S. Park, "A study of speech emotion recognition and its application to mobile services," in *Proc. of the International Conference on Ubiquitous Intelligence and Computing*. Springer, 2007, pp. 758–766.

[4] P. Gupta and N. Rajput, "Two-stream emotion recognition for call center monitoring," in *Proc. of the 8th Annual Conference of the International Speech Communication Association INTERSPEECH*. ISCA, 2007, pp. 2241–2244.

[5] M. Szwoch and W. Szwoch, "Emotion recognition for affect aware video games," in *Image Processing & Communications Challenges 6*, R. S. Choraś, Ed. Springer International Publishing, 2015, pp. 227–236.

[6] D. Van Lancker, C. Cornelius, and J. Kreiman, "Recognition of emotional-prosodic meanings in speech by autistic, schizophrenic, and normal children," *Developmental Neuropsychology*, vol. 5, no. 2-3, pp. 207–226, 1989.

[7] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574–586, 2010.

[8] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[9] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[10] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, 2012.

[11] S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin, "A review on emotion recognition using speech," in *Proc. of the ICICCT*. IEEE, 2017, pp. 109–114.

[12] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 7–19, 2010.

[13] L. Tian, J. Moore, and C. Lai, "Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features," in *Proc. of the IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 565–572.

[14] H. Kaya, D. Fedotov, A. Yesilkanat, O. Verkholyak, Y. Zhang, and A. Karpov, "LSTM based cross-corpus and cross-task acoustic emotion recognition." in *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Hyderabad, India: ISCA, 2018, pp. 521–525.

[15] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. of IITERSPEECH 2014*. ISCA, 2014, pp. 223–227.

[16] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[17] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Towards speech emotion recognition" in the wild" using aggregated corpora and deep multi-task learning," in *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. ISCA, 2017, pp. 1113–1117.

[18] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[19] D. Le, Z. Aldeneh, and E. M. Provost, "Discretized continuous speech emotion recognition with multi-task deep recurrent neural network." in *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017, pp. 1108–1112.

[20] X. Ouyang, S. Kawaai, E. G. H. Goh, S. Shen, W. Ding, H. Ming, and D.-Y. Huang, "Audio-visual emotion recognition using deep transfer learning and multiple temporal models," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 577–582.

[21] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3030–3043, 2017.

[22] I. Kansizoglou, L. Bampis, and A. Gasteratos, "An active learning paradigm for online audio-visual emotion recognition," *IEEE Transactions on Affective Computing*, 2019.

[23] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-visual emotion fusion (avef): A deep efficient weighted approach," *Information Fusion*, vol. 46, pp. 184–192, 2019.

[24] Z. Wang, L. Wang, and H. Huang, "Joint low rank embedded multiple features learning for audio–visual emotion recognition," *Neurocomputing*, vol. 388, pp. 324–333, 2020.

[25] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-visual emotion recognition," *Pattern Recognition Letters*, 2021.

[26] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks," *Synthesis Lectures on Computer Architecture*, vol. 15, no. 2, pp. 1–341, 2020.

[27] P. Viola, M. Jones *et al.*, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 4, no. 4, pp. 34–47, 2001.

[28] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[29] J. F. Gemmeke, D. P. 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. of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[30] P. Ekman, "An argument for basic emotions," *Cognitive Emotion*, vol. 6, pp. 169–200, 1992.

[31] A. Shilandari, H. Marvi, H. Khosravi, and W. Wang, "Speech emotion recognition using data augmentation method by cycle-generative adversarial networks," *Signal, Image and Video Processing*, 2022.

[32] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011.

[33] D. Hu, X. Li *et al.*, "Temporal multimodal learning in audiovisual speech recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3574–3582.

[34] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1583–1597, 2016.

[35] L. C. De Silva and P. C. Ng, "Bimodal emotion recognition," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. IEEE, 2000, pp. 332–335.

[36] L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu, "Multimodal human emotion/expression recognition," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 1998, pp. 366–371.

[37] L. Huang, L. Xin, L. Zhao, and J. Tao, "Combining audio and video by dominance in bimodal emotion recognition," in *Proc. of the International Conference on Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 729–730.

[38] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. of the 6th International Conference on Multimodal Interfaces*. ACM, 2004, pp. 205–211.

[39] Y. Yoshitomi, S.-I. Kim, T. Kawano, and T. Kilazoe, "Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face," in *Proc. of the 9th IEEE International Workshop on Robot and Human Interactive Communication*. IEEE, 2000, pp. 178–183.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[41] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.

[42] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK*, 2014.

[43] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. 60, pp. 1755–1758, 2009.

[44] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[45] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[46] S. Zhao, Y. Ma, Y. Gu, J. Yang, T. Xing, P. Xu, R. Hu, H. Chai, and K. Keutzer, "An end-to-end visual-audio attention network for emotion recognition in user-generated videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 303–311.

[47] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[48] S. Zhao, Y. Ma, Y. Gu, J. Yang, T. Xing, P. Xu, R. Hu, H. Chai, and K. Keutzer, "An end-to-end visual-audio attention network for emotion recognition in user-generated videos," in *In Proc. of AAAI*, 2020.