Multimodal Emotion Recognition from Voice and Video Signals

Conference Paper · August 2023

DOI: 10.1109/EUROCON56442.2023.10198928

3 authors, including:

Danilo Greco
Università degli Studi di Genova
12 PUBLICATIONS **17** CITATIONS

Multimodal Emotion Recognition from Voice and Video Signals

Paola Barra
DIST
University of Naples Parthenope
Naples, Italy
paola.barra@uniparthenope.it

Zied Mnasri
DAAM
University of Naples L'Orientale
Naples, Italy
zmnasri@unior.it

Danilo Greco
DiSEGIM
University of Naples Parthenope
Naples, Italy
danilo.greco@uniparthenope.it

IEEE EUROCON 2023 - 20th International Conference on Smart Technologies | 978-1-6654-6397-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/EUROCON56442.2023.10198928

***Abstract*—A promising area of research and development that can significantly increase the efficacy and accuracy of mental health assessments is the use of artificial intelligence (AI) and machine learning algorithms to analyse simultaneously voice and facial expressions in a video stream. More studies are required to completely comprehend the capabilities and limitations of these technologies and guarantee their ethical and effective usage in clinical settings. Collaborative robots (cobots) have the potential to completely change how mental evaluations of autistic children are approached. ChatGPT is an effective language model that can understand and produce human-like text. When used in conjunction with the Cobot, this technology enables children with autism to interact and communicate in a way that is natural to them. In this article, we introduce a novel method for analysing emotional detection using voice analysis and facial recognition that has been tested on the IEMOCAP database. The outcomes session, which illustrates the tool's potential use in healthcare, concludes the paper.**

***Index Terms*—Cobot, emotion recognition, spoken language, facial expressions, deep learning, convolutional neural network (CNN), transformers, ChatGPT, deep learning model types, spectrogram image, web-shaped model (WSM)**

I. Introduction

In the field of digital health, there has been extensive research on emotion recognition utilising spoken language and facial expressions [1], [2]. Deep learning techniques are also included among the strategies involved to solve this problem [3]: it is possible to train convolutional neural networks (CNN) on large collections of photos with the proper emotions labelled on them and subsequently recognize picture patterns associated with emotions [4]. Furthermore recurrent neural networks (RNN) can be used to analyze audio data for spoken words [5]. An assortment of audio recordings that have each been given a different emotion can be used to train the RNN. The network can then develop the ability to identify audio data patterns that represent various emotional states. The CNN and RNN can both be trained independently before being integrated into a single model that can identify emotions through both spoken and facial expressions. To distinguish emotions from spoken language and facial expressions, other deep learning model types, such as long short-term memory (LSTM) networks or transformers, can also be utilised [6]. The decision around the model to be employed will depend on the specific characteristics of the data. In this paper, we present a novel approach [7], [8] to recognize emotions from speech and images coming from a video data set. The simultaneous analysis of voice and facial expressions in a video stream presents a promising field to improve the efficacy and accuracy of mental health assessments. The experiments were performed on the IEMOCAP[1] dataset which provides video and audio labelled by time interval with related emotions. More research is necessary to fully understand the potential and constraints of these technologies [9], as well as to ensure their effective and moral application in therapeutic contexts. Compared with more conventional robotic systems, Cobots, or collaborative robots, have a number of advantages as: they can be configured to communicate with children in a fun and relational way; they can be trained to interact with children in interesting and reassuring ways; cobots can do things such as point to objects or make certain movements, which can help engage children and provide a more thorough assessment of their abilities. The way mental assessments of autistic children are performed could change completely through the interaction between ChatGPT and Cobots. The use of these cutting-edge technologies could offer a more effective and efficient way to assess and understand the needs of these children, A system that integrates language and picture analysis with patient contact with robots to evaluate their social and communication skills might be developed to research autistic pathophysiology. The system might engage with patients (children) via voice and facial-recognition-equipped Cobots while gathering data on their feelings and behaviours. In order to spot any irregularities in communication and emotional expression, the system could also evaluate the voices and facial expressions of the patients. The severity of the condition might be determined using this research data, and more individualized and focused therapy interventions could be

[1] https://sail.usc.edu/iemocap/

978-1-6654-6397-3/23/$31.00 ©2023 IEEE

created.

II. Related work

Audio-visual emotion recognition has been recently reviewed in several survey publications, e.g. [10], that examine experimental methods for multimodal emotion recognition. Methods used for audio-visual emotion recognition can be categorized into three main families: a) generative methods, b) supervised learning and c) unsupervised learning.

A. Generative modeling

Mainly using Bayesian models, where each emotion category is modelled by a GMM/HMM. For instance, in order to fuse numerous signals for an Error-Weighted Classifier (EWC), [11] adopted a Bayesian framework to blend empirical evidence with previous assumptions. An HMM was trained for each emotional phonetic category in the voice modality. In the face modality, the GMM trained for each emotion without having access to visual information models the top face, while HMMs trained for each emotional visual information model the bottom face. In order to properly integrate different modalities, the weighted total of the individual judgments was merged after investigating their contributions, which were determined using the confusion matrices of each classifier.

B. Supervised learning

Supervised learning is definitely the most used for several years for emotion recognition in conversational speech, either from audio or video. For instance, [10] reported that:

- By taking into account the speaker information of the target utterance and additionally modelling self and inter-speaker emotional effect with a hierarchical multi-stage RNN with attention mechanism, DialogueRNN [12] seeks to resolve this issue.
- By utilizing quantum theory and RNN-LSTM, the necessity to understand inter-speaker interdependence for emotion recognition in conversational speech is addressed and modelled in [13].
- Interaction-Aware Attention Network (IANN), an emotion recognition technique for conversational speech based on inter-speaker connection modelling, was recently proposed by [14], where each speaker is modelled by a unique memory.

C. Unupervised learning

Unsupervised learning can be useful either for emotion clustering or feature extraction. Emotion clustering consists in discovering emotion groups which do not necessarily match with unique labels. For e.g. a cluster grouping 'excitement' and 'anger' signals may indicate 'high arousal', whereas 'calm' and 'sadness', when clustered together, indicate 'low arousal'. Then, such a type of clustering would be useful in proposing a novel classification of emotions. This track has been investigated in [15], where fuzzy clustering helped construct a membership matrix that shows that different emotion categories may share similar valence or arousal characteristics. On the other hand, unsupervised autoencoders were used by [16] to extract latent features for speech emotion recognition. An autoencoder is a neural network that approximates the identity function. However, its interest for this application consists in collecting latent features at the code layer, which may be more useful than hand-crafted features to characterize speech for the particular task of comparing the emotional content of speech.

III. Materials and methods

A. Emotion recognition datasets

They can be classified into three main categories: a) speech emotion datasets, b) facial expression datasets and c) video and/or multimodal emotion recognition. Most of these datasets use either spontaneous or acted speech or scenes. The corresponding labels are generally provided either in a categorical or a dimensional scheme. Categorical labels usually follow either the basic emotion model proposed by Eckman or Plutchik's compound model, known also as the wheel of emotions. On the other hand, the dimensional model represents emotions in a 3-D coordinate system where each emotion is characterized by a score of i) valence, i.e. positive or negative, ii) activation, i.e. high or low, and iii) dominance [15], as illustrated in Fig. 1.

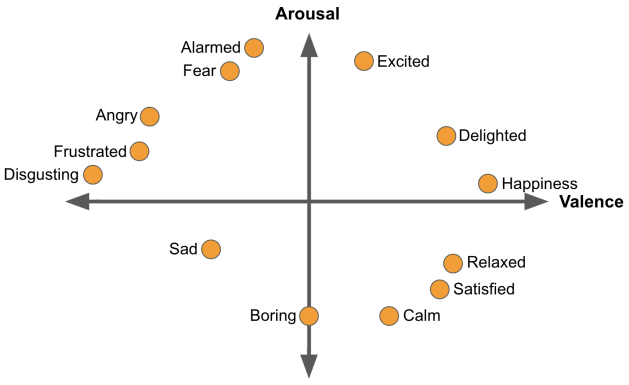

Fig. 1: Emotion categories and dimensional attributes.

Therefore, most emotion recognition datasets are built following either the basic, the compound or the dimensional model. For instance, speech emotion recognition datasets like EMO-DB and EMOVO contain every 6 basic emotions plus neutrality [10]. However, multimodal datasets like IEMOCAP [17], CMU-MOSEI, MELD and SEMAINE [10] contain a more complex labelling system, including both basic emotion categorical labels or scores for the dimensional axes, either for improvised or acted scenes. For example, in IEMOCAP dataset, categorical labels and dimensional scores are given by several human evaluators, to yield a majority voting. More details about

state-of-the-art datasets used in speech and multimodal emotion recognition can be found in [10].

B. Speech emotion recognition

In recent years, a substantial body of research has been conducted on the use of convolutional neural networks (CNNs) for voice emotion recognition. The complexity and dynamic nature of speech signals, which are influenced by several elements like speaking style, pronunciation, and accent, make it difficult to recognize emotions in speech. However, it is also a vital area of research since emotions are fundamental to communication and their understanding can have a big impact on a variety of applications, including speech-based human-computer interaction, affective computing, and mental health diagnostics.

1) Low-level descriptors: They represent a set of features intentionally designed for speech emotion recognition. Such features have been selected to form the well-known GeMAPS feature set [18]. This feature set includes several types of prosodic, acoustic and spectral descriptors, computed at the raw signal level. Several variants of GeMAPS features are available in OpenSmile toolbox [19], each for a particular task in affective speech computing. These features were proven to give high affective speech recognition rates in several challenges, such as Interspeech'09 emotion challenge [20], Computational Paralinguistics challenge (ComParE) [21]. However, they belong to the paradigm of explicit feature extraction, whereas novel feature extraction methods are moving towards the concept of an end-to-end learning process.

2) Feature extraction and classification methods: A more recent technique of feature extraction in speech emotion recognition consists in the usage of spectrogram images as an implicit representation of the Mel-Frequency Cepstral Coefficients (MFCCs) [22]. A spectrogram is a time-frequency representation of speech signals that can be used to identify the spectral and temporal aspects of speech, which are significant markers of emotional state. The collection of MFCCs, on the other hand, is resilient to changes in speaking style, pronunciation, and accent. They summarize the spectral envelope of speech signals. The spectrogram can be utilized as input for speech emotion classification in several ways by a) extracting feature vectors composed of MFCC coefficients and their Δ and Δ-Δ, i.e. their first and second derivatives; b) feature embedding, usually through learning latent features collected at the code layer of an autoencoder [15], or c) presenting the spectrogram image as raw input to the classifier.

When it comes to using spectrogram images as input, convolutional neural networks (CNN) are preferred as classifier methods, since they excel in classifying images and they can accurately identify patterns in the input data. When recognizing emotions in speech, a CNN uses the spectrogram or MFCCs as input and trains on a sizable dataset of speech signals to learn how to categorize the emotions. The combination of feature extraction, CNNs, and numerous other techniques can considerably enhance the precision, robustness, and generalization of emotion recognition in speech. The absence of significant and diverse datasets, the difficulty of defining a set of emotions that is consistent, and the impact of numerous confounding variables on speech signals are just a few of the difficulties the area is now experiencing.

C. Facial emotion recognition

From learning platforms to human-computer and human-robot interaction, facial emotion recognition (FER) is widely used in a variety of activities. It can also be used as an intrinsic technique for face recognition issues to produce an expression-free face classifier. Most approaches focus on building ever-deeper neural networks that regard an expression as a still image or, in certain cases, as a sequence of succeeding frames that reflect the expression's temporal component [24].

1) Feature extraction and classification methods.: FER usually starts with the extraction of facial features, which is strictly succeeded by an emotion classification method. This section presents the most recent works that led to the experimentation of different techniques. Methods for feature extraction are mainly divided into geometric methods and appearance-based methods. In particular, geometric methods focus on the shape, scale, orientation, and location of the various parts of the face, such as the nose, mouth, eyes and eyebrows [25]. In this context, Active Shape Model is a feature-matching solution that focuses on point features and measures several shape variations and a range of adaptive models on the face [26]. The authors in [27] propose a technique that consists of an estimate of parameters extracted from the wavelet transform, and then an SVM classifier performs the emotion classification. In [28], oriented gradient histograms (HOG), Gabor, and local binary pattern (LBP) have been used as feature extractors; subsequently, a simple k-nearest neighbour (kNN) was used for the classification. The experiments carried out aimed to demonstrate the effectiveness of these three feature extractors in the FER. The authors in [29] propose a system of FER jointly with the gender and age of the subject to understand how much these affect facial expressions. The authors used the classic KNN and SVM classifiers and deep learning models such as CNN and VGG16.

2) Web Shaped Model: The computational cost of the training phase applies to these, which could take hours or days to complete. This work proposes the Web Shaped Model (WSM), a geometric method to extract features that can distinguish between various facial emotions. It uses a virtual geometric pattern resembling a spider's web drawn on a face. The Web Shaped model was introduced in [7] to detect the pose of the face; later, it was also used for the FER [8]. This method: (1) locates face landmarks with the Kazemi-Sullivan method [30]; (2) draws the web centred on the nose landmark; (3) keeps the web with

the associated emotion tag. The resulting coding is an array of how many landmarks fall into each sector of the web. It locates facial regions containing facial landmarks. Fig. 2 shows an example of this process for an emotion 'neutral' image. The resulting encoding varies depending on the choice of the number of concentric circles and radii; the final array changes its size and content. The web configuration used has 60 rays and 8 concentric circles, resulting in 480 elements in the final coding

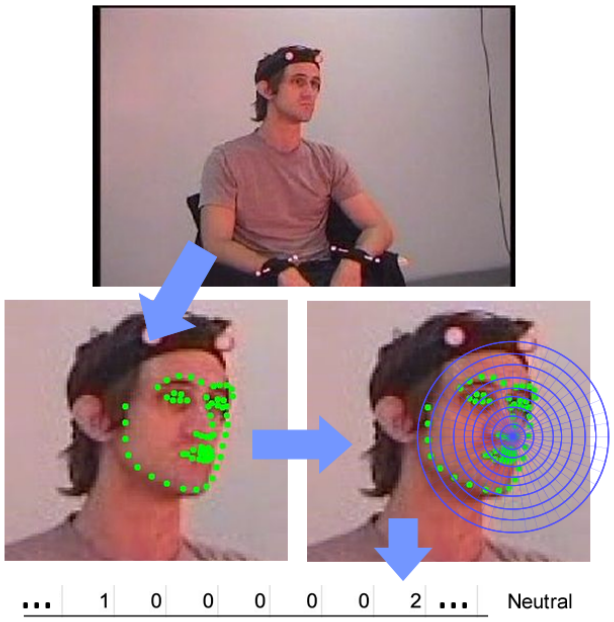

Fig. 2: Process of creating emotion coding using WSM.

IV. Experiments and results

A. Experimental protocol

An experimental setup has been established to carry out both tasks simultaneously and jointly in order to evaluate the efficacy of merging spoken language and facial expressions for emotion recognition. As a result, the following tests were carried out using conversational speech videos that were either improvised or scripted and were included in the IEMOCAP dataset [17]:

- Conversational video sessions were segmented into chunks, each containing only one sentence, uttered by a single speaker.
- For each frame, a categorical label has been defined among the following basic emotions: *happiness, anger, sadness, fear, neutral, other*. It should be noted that some class labels provided in the dataset were merged with other ones for their reduced occurrence, e.g excitement and surprise are labelled as happiness and frustration as sadness.
- Each audio chunk is segmented into short frames (audio frames) of 100 ms each with 75% overlap.
- For each audio frame, a series of spectrograms are extracted to be trained as input features.
- The video of each sentence is segmented into successive images.
- For each audio frame or image, the corresponding label is that of the sentence.
- Either for audio frames or images taken from each video, feature extraction and training are performed separately, using different types of features and classifiers for each task.
- After classifying the single audio frames and images, a majority voting is applied to predict the label at the sentence level.
- The performance of each classifier is evaluated using typical metrics for supervised learning, i.e. overall accuracy, class-wise precision, recall and F1-score.

The combined process is illustrated in Fig. 3, both for facial images and speech emotion classification.

B. Audio classification for speech emotion recognition

To implement the audio signal classifier, we opted to use spectrogram images as input and CNNs as classifiers. Thus, no explicit features, such as OpenSmile LLD's [19] were used. The rationale behind such a choice is simply empirical, based on the performance of each type of input. Feature extraction and classification are achieved at the frame level, using the following process:

- Each audio signal, generally corresponding to one uttered sentence, is segmented into overlapping short frames, of 100 ms each. Actually, such a duration is the minimum that can carry emotion.
- For each frame, a series of spectrogram images is extracted, applying a short-time Fourier transform at overlapping windows.
- The spectrogram images are fed into a CNN network, composed of several 2-D convolutional layers, each followed by a max-pooling layer. The last layer is a 1-D Dense layer with a softmax activation that returns a classification probability.
- Once all frames belonging to a given sentence are classified, a majority voting is applied to decide about the sentence's label.

The results of this workflow obtained on a separate test set randomly extracted from the same session of IEMOCAP database used for training are mentioned in Table I.

C. Image classification for face emotion recognition

For each time interval the images, the WSM coding from each face was extracted and labelled with the corresponding emotion. The classifier was then trained and tested with the same fragments used in audio testing. In previous articles, the spider web has been used only for a single face, in this work the emotions have been classified by video intervals. Each image in the video range is classified and the classification is carried out for a single frame.

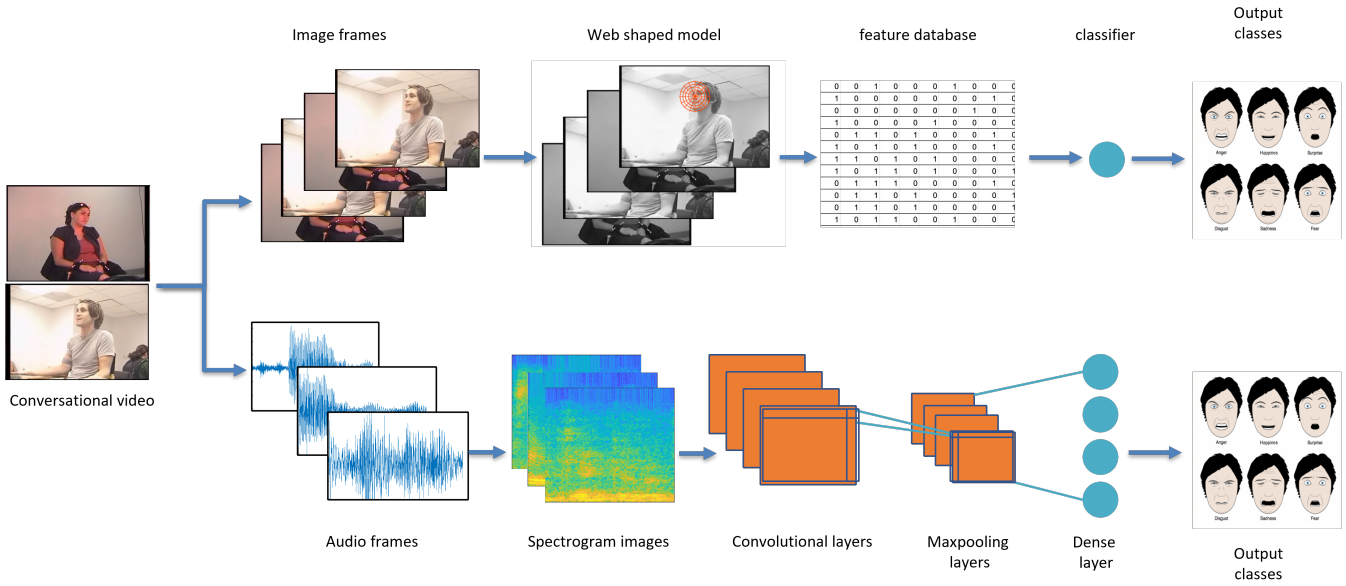

Fig. 3: Workflow of the combined emotion recognition process from facial images and speech signal

TABLE I: Emotion classification results from voice signals and facial images

Emotion Class	Voice signals				Facial images			
	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
Anger	0.76	0.30	0.43		0.15	0.47	0.23	
Fear	1.00	0.17	0.29		0.33	0.50	0.40	
Happiness	0.77	0.47	0.58		0.81	0.60	0.69	
Sadness	0.58	0.95	0.72		0.65	0.68	0.66	
Neutral	1.00	0.21	0.35		0.63	0.33	0.44	
Other	1.00	0.25	0.4		0.00	0.00	0.00	
All				0.63				0.60

Of all the predicted classes in the range, the majority voting value is considered to determine the predicted class in the whole fragment. The classification was done with various known state-of-the-art classifiers to compare with which we obtained the best performance. The best performing classifier is the Random Forest Classifier with the parameters:

- The whole dataset is used to build each tree.
- tree splitting criterion = entropy. This criterion is equivalent to minimizing the cross-entropy and multinomial deviance) between the true labels and the probabilistic prediction.
- number of estimators = 1000.

D. Discussion

Results show differences for each type of input, i.e voice signals and facial images, and for each category of emotion. Whereas some emotions, like *happiness* and *sadness* are pretty well recognized, either by audio or image classification, other ones are much less recognized. For instance, audio-based classification presents low recall rates for most emotions, which indicates high false positive rates. This may be due to the inability of the spectrogram to decorrelate emotions and voices. Actually, once the CNN is trained on a certain pattern of spectrogram, it learns the characteristic spectral coefficients of the voice, and hence, becomes more prone to recognize the voice than the emotion. In other words, if a voice is frequently labelled as *happy*, it would also be labelled so even if the target emotion is different. FER results are highly dependent on the type of images acquired. The emotions of the IEMOCAP dataset are realistic and include many microexpressions that are unlikely to be captured by cameras with a framerate lower than 30FPS [32]. The problem in question is of a technical nature and also justifies the low result of the *other* class, which is a collection of unidentified emotions. Otherwise, the general classification trend follows speech recognition in which classic emotions such as *sadness* and *happiness* are quite well classified.

V. Conclusions

Multimodal emotion recognition from audio and video channels can open new avenues in mental health monitoring. Information gathered from the conversation with a Cobot and ChatGPT can be reviewed in real-time, providing the evaluator with immediate feedback. Since children with autism often have difficulty taking typical exams that rely on lengthy questionnaires or standardized exams, this

can be extremely useful for them. To explore this avenue, we presented preliminary results on emotion recognition on the IEMOCAP database, which provides voice data and video. Experiments demonstrate how separately it is possible to detect this information. Future experiments may consist in: a) improving the results of emotion recognition by each type of classifier, b) integrating both classifiers into a single multimodal framework obtained from the conjunction of both extracted feature sets using data integration with fuzzy methods [31], and applying visual transformers to introduce recurrent learning.

References

[1] P. Harár, R. Burget and M. K. Dutta, "Speech emotion recognition with deep learning," 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2017, pp. 137-140, doi: 10.1109/SPIN.2017.8049931.

[2] Hadhami Aouani, Yassine Ben Ayed, "Speech Emotion Recognition with deep learning", Procedia Computer Science, Volume 176, 2020, Pages 251-260, ISSN 1877-0509.

[3] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.

[4] B. Zhang, C. Quan and F. Ren, "Study on CNN in the recognition of emotion in audio and images," 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Okayama, Japan, 2016, pp. 1-5, doi: 10.1109/ICIS.2016.7550778.

[5] Graves, Alex. "Generating sequences with recurrent neural networks." In arXiv preprint arXiv:1308.0850, 2013.

[6] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. Neural Comput. 9, 8 (November 15, 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[7] P. Barra, S. Barra, C. Bisogni, M. De Marsico and M. Nappi, "Web-Shaped Model for Head Pose Estimation: An Approach for Best Exemplar Selection," in IEEE Transactions on Image Processing, vol. 29, pp. 5457-5468, 2020, doi: 10.1109/TIP.2020.2984373

[8] Barra, P., De Maio, L. and Barra, S. Emotion recognition by web-shaped model. Multimed Tools Appl (2022). https://doi.org/10.1007/s11042-022-13361-6

[9] Z. Tariq, S. K. Shah and Y. Lee, "Speech Emotion Detection using IoT based Deep Learning for Health Care," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 4191-4196, doi: 10.1109/BigData47090.2019.9005638.

[10] S. Poria, N. Majumder, R. Mihalcea and E. Hovy, "Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances," in IEEE Access, vol. 7, pp. 100943-100953, 2019, doi: 10.1109/ACCESS.2019.2929050.

[11] A. Metallinou, S. Lee and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 2010, pp. 2462-2465, doi: 10.1109/ICASSP.2010.5494890.

[12] Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., and Cambria, E. (2019). DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 6818-6825. https://doi.org/10.1609/aaai.v33i01.33016818

[13] Y. Zhang, Q. Li, D. Song, P. Zhang, and P. Wang, "Quantum-inspired interactive networks for conversational sentiment analysis", in Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI), 2019, pp. 1–8.

[14] Yeh, Sung-Lin, Yun-Shao Lin, and Chi-Chun Lee. "An interaction-aware attention network for speech emotion recognition in spoken dialogs." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[15] Rovetta, Stefano, et al. "Emotion Recognition from Speech: An Unsupervised Learning Approach." Int. J. Comput. Intell. Syst. 14.1 (2021): 23-35.

[16] Rovetta, Stefano, Zied Mnasri, and Francesco Masulli. "Emotional content comparison in speech signal using feature embedding." Progresses in Artificial Intelligence and Neural Systems (2021): 45-55.

[17] Busso, Carlos, et al. "IEMOCAP: Interactive emotional dyadic motion capture database." Language resources and evaluation 42 (2008): 335-359.

[18] Eyben, Florian, et al. "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing." IEEE transactions on affective computing 7.2 (2015): 190-202.

[19] Eyben, Florian, and Björn Schuller. "openSMILE:) The Munich open-source large-scale multimedia feature extractor." ACM SIGMultimedia Records 6.4 (2015): 4-13.

[20] Schuller, Björn, Stefan Steidl, and Anton Batliner. "The interspeech 2009 emotion challenge." (2009).

[21] Schuller, Björn, et al. "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language." 17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5. Vol. 8. ISCA, 2016.

[22] R. Vergin, D. O'Shaughnessy and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," in IEEE Transactions on Speech and Audio Processing, vol. 7, no. 5, pp. 525-532, Sept. 1999, doi: 10.1109/89.784104.

[23] Scirea, Marco, et al. "Mood expression in real-time computer-generated music using pure data." Proceedings of the ICMPC-APSCOM 2014 Joint Conference. College of Music, Yonsei University, 2014.

[24] Lee, Sunny, A Brief Review of Deep Learning for Facial Expression Recognition (January 5, 2023). Available at SSRN: https://ssrn.com/abstract=4318896 or http://dx.doi.org/10.2139/ssrn.4318896

[25] Navjot Rathour, Rajesh Singh, Anita Gehlot, Shaik Vaseem Akram, Amit Kumar Thakur, Amit Kumar, The decadal perspective of facial emotion processing and Recognition: A survey, Displays, Volume 75, 2022, 102330, ISSN 0141-9382, https://doi.org/10.1016/j.displa.2022.102330.

[26] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active Shape Models-Their Training and Application, Computer Vision and Image Understanding, Volume 61, Issue 1, 1995, Pages 38-59, ISSN 1077-3142, https://doi.org/10.1006/cviu.1995.1004.

[27] Jeen Retna Kumar, R., Sundaram, M., Arumugam, N. et al. Face feature extraction fohr emotion recognition using statistical parameters from subband selective multilevel stationary biorthogonal wavelet transform. Soft Comput 25, 5483–5501 (2021). https://doi.org/10.1007/s00500-020-05550-y

[28] Subudhiray, Swapna, Hemanta Kumar Palo, and Niva Das. "K-nearest neighbor based facial emotion recognition using effective features." IAES International Journal of Artificial Intelligence 12.1 (2023): 57.

[29] Surya Teja Chavali, Charan Tej Kandavalli, T M Sugash, R Subramani, Smart Facial Emotion Recognition With Gender and Age Factor Estimation, Procedia Computer Science, Volume 218, 2023, Pages 113-123, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2022.12.407.

[30] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 1867-1874, doi: 10.1109/CVPR.2014.241.

[31] Ciaramella, A., Nardone, D. and Staiano, A. Data integration by fuzzy similarity-based hierarchical clustering. BMC Bioinformatics 21 (Suppl 10), 350 (2020). https://doi.org/10.1186/s12859-020-03567-6

[32] Borza D, Danescu R, Itu R, Darabant A. High-Speed Video System for Micro-Expression Detection and Recognition. Sensors (Basel). 2017 Dec 14;17(12):2913. doi: 10.3390/s17122913. PMID: 29240700; PMCID: PMC5751645.