Speech Emotion Recognition Using Deep Learning

Jennifer C Saldanha[1] and Rohan Pinto[2]
[1,2]*Department of Electronics and Communication Engineering,*
St Joseph Engineering College, Mangaluru
Affiliated to Visvesvaraya Technological University, Belagavi

***Abstract*—Speech emotion recognition a space-growing analysis domain in recent years. Unlike humans, machines lack the skills to understand and show emotions, however, human-machine interactions are often improved by automatic emotion recognition, thereby reducing the necessity of human intervention. An SER system is a group of techniques for classifying and processing speech signals in order to find any embedded emotions. In this work, the RAVDEES database for speech emotion recognition is selected from Kaggle. The MFCC feature is extracted. Deep learning algorithm, CNN is used which classifies the extracted relevant MFCC features of speech signals which are used and recognizes the emotion. The speech emotion recognition system eases the identification of the speaker's emotion and mental status. CNN model implemented in this work can recognize the emotional state of the speaker. The project achieved training accuracy of 96% and testing accuracy of 85%. This results in an accurate identification of the emotion.**

***Index Terms*—Speech Emotion Recognition, Mel Frequency Cepstral Coefficients, Convolutional Neural Network, Long Short Time Memory, Deep Belief Network, Recurrent Neural Network.**

I. INTRODUCTION

Speech is a common and natural way for people to communicate with one another. Emotion Recognition examines how emotions are inferred and the techniques used to do the recognition [1]. Speech patterns and facial expressions both show emotion. Emotion detection from voice has developed from a niche application to a critical component of Human-Computer Interaction (HCI). By using direct voice contact as input to understand verbal information and make it simple for human listeners to respond, these systems attempt to facilitate the natural engagement with machines. In order to detect the dissatisfied customer, customer satisfaction, and other factors, conversational analysis uses emotion recognition as a performance parameter [2]. On-board driving systems, call center chats, dialogue systems for spoken languages, and the use of emotional speech patterns in medical applications are a few examples. However, there are a few issues with HCI systems that should also be properly addressed, especially when these systems go from being tested in research labs to being put into use. Therefore, efforts are needed to successfully address these issues and raise the level of robots' ability to recognize emotions. People around the world have different cultural origins, native tongues, speaking tempos, and speaking styles. Due to these cultural distinctions, it is more challenging to correctly determine the speaker's emotional states and to select the appropriate speech patterns.

The attributes utilized in recognition of emotions were resulting from alterations in facial mimics similarly as speech indications. Feeling remains as a physiological response that ensues in things like sadness, worry or happiness. There are few more emotions like Neutral, Anger, disappointment within which any intelligent system with finite machine resources may be trained to spot or synthesize as required [1]. In this work, a deep learning technique is adapted that automatically extracts the relevant emotional features of the speech signal. These neural networks enable learning from vast volumes of data in an effort to mimic the functioning of the human brain. Additional hidden layers can help to optimize and improve for accuracy even though a single-layer neural network may only give approximate predictions. Some of the data pre-processing that machine learning typically requires is eliminated by deep learning [3]. By automating feature extraction and ingesting and interpreting unstructured data, including text and images, these algorithms eliminate the need for human experts.

II. LITERATURE REVIEW

There are numerous other recent publications and surveys on SER due to its significance in human-

computer interaction and the advancement of artificial intelligence systems. The most recent studies that are connected to the subject are reviewed in this section.

The paper [4] discusses about the speech emotion recognition (SER) using Pattern Recognition Neural Network (PRNN), K – Nearest Neighbor (KNN) and Gray Level Co-occurrence Matrix (GLCM) [5] which has been used as a primary tool in image texture analysis. Using well-known speech emotion identification approaches, such as the Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM), the results acquired in this study were assessed for precision rate, F-Measure and accuracy and were shown to produce significant results.

In order to increase the accuracy of speech emotion recognition, the authors [6] proposed a novel speech emotion recognition method based on Gaussian Kernel Nonlinear Proximal Support Vector Machine (PSVM) in the classifying the emotions such as angry, joy, sadness and surprise. First, perform preprocessing on the speech signal, including sampling, quantification, pre-emphasizing, framing, window addition, and endpoint detection. Second, take note of speech prosody and qualities. SER method is proposed using a CNN model. It is known that constructing a neural network with MTL requires the model to share hidden layers in front of the output decision layers for all main and auxiliary tasks. In addition, careful selection of useful subtasks is very important, since unrelated subtasks may degrade performance [6].

The performance of emotion recognition system was evaluated by considering EMO-DB dataset [7]. In [8] performance of random forest classifier is evaluated on the emotion recognition application using speech samples. Gini impurity was used to gauge how well a node was split and chose to separate nodes until each leaf had samples from only one class after testing various maximum depths. The maximum accuracy of 81.05% was attained using hyper-parameter optimization techniques.

In [9] analyzed the significance of several classification algorithms, including SVM and HMM, after publishing a brief analysis on the significance of speech emotion datasets, features and effect of noise reduction on the recognition task. The research's strength is the discovery of a number of features associated with voice emotion recognition. The usage of CNN and Recurrent Neural Network (RNN) was explored as a deep learning technique. Deep learning algorithms, such Deep CNNs, have recently been demonstrated to be able to learn emotional aspects for SER in [10]. Multitask learning based on a CNN was used for SER. It involves arousal level, valence level, and categorization based on gender. As a result, it can offer more accuracy and lessen the generalizing error issue. With the FAU Aibo dataset, a frame-level hybrid deep belief network (DBN) and HMM classifier was suggested [3] to categorize the five emotion classes. It was suggested to use a recurrent neural network (RNN) to implement an automated SER system. In this work a bidirectional LSTM with a pooling method that concentrates on the emotionally significant portions of utterances was used. In [11] the IEMOCAP dataset was used for the performance evaluation of the classification algorithm. CNN model was put into practice using Tensor Flow. Over 400 spectrograms were extracted from speech signal. There were around 500 images for each category of emotion. The procedure of training was run with a batch size of 100 for a period of 20 epochs. Recurrent Neural Network (RNN), Deep Neural Network (DNN), and spectral feature extraction have been applied on different datasets in [12] and recursive feature elimination (RFE) model has been used for feature selection.

III. METHODOLOGY

The data is split into training and testing data. The next step is to separate the valuable features from the gathered data so they can be fed into the model as input. The model must then be assembled, fitted with training data, and evaluated using a test dataset. Predictions and classification will be made using the model when it has been trained and tested. Fig. 1 shows the block diagram of emotion recognition system.

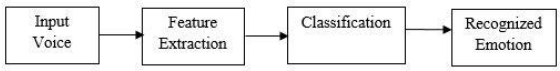

Fig. 1. Block diagram of the speech emotion recognition

A. Input Voice

The input for SER is a speech signal from which emotions must be identified. The emotional database samples are used as input sources [13]. RAVDEES (Ryerson Audio-Visual Database of Emotional Speech

and Song) dataset is used as input source for SER. The RAVDESS, has 1440 files from 24 actors. Totally 60 trials are taken from each actors. The 24 professional actors (12 male and 12 female) perform two lexically similar phrases in the RAVDESS with a neutral North American accent. Speech emotions composed of expressions of calmness, joy, sadness, anger, fear, surprise, and disgust. There are two emotional intensity levels and one neutral expression created for each expression.

B. Feature Extraction

The process of extracting features by the requirements is known as feature extraction. In this work, MFCC is used for feature extraction method as it is expected to provide accurate results due to its ability in automatic recognition of speech [14]. The final accuracy is determined by the feature quality. In the training and testing phases, the acquired signals is used. A representation of a sound's short-term power spectrum used in sound processing is called a Mel-frequency cepstrum. It is based on a linear cosine transform of a log-range spectrum on a nonlinear Mel frequency scale. Signal extraction frequently makes use of the signal's Fourier transform. Then, transfer the obtained spectrum's powers onto the Mel scale using triangular overlapping windows. After discrete cosine modifications, log the powers at each Mel frequency. The sound signal is represented by the amplitude as an MFCC. The Mel-frequency cepstrum (MFC) is based on a linear cosine conversion of a log power spectrum on a nonlinear Mel frequency scale. It is a representation of a sound's short-term power spectrum. An MFC is made up of coefficients known as Mel-frequency cepstral coefficients (MFCCs). The DFT, the log of the magnitude calculation, windowing the signal, warping the frequencies on a Mel scale, and inverse DCT are all steps in the MFCC feature extraction process. MFCC is the most important and useful method in speech-related applications. The vocal tract's anatomy filters the sounds that people make, producing only certain sounds. The short-time power spectrum envelope shows the vocal tract. The linear cosine transform of a log power spectrum is used by MFCC to depict the short-term power spectrum of sound. The output obtained from MFCC feature is numerical value, in this work since CNN is used for classification, the input should be in the form of a 2D image. Hence the obtained numerical values are formed as a 2D array and it is given as input to CNN. Fig. 2 shows MFCC features and visualization plot obtained as simulation output using MATLAB software.

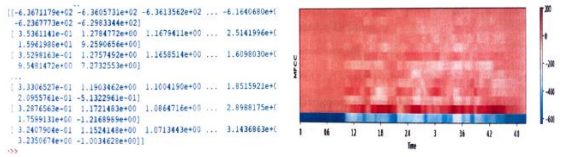

Fig. 2. MFCC Features and Visualization

C. Classification

The CNN algorithm of deep learning, which consists of convolutional layers, max-pooling layers, and fully linked layers, is used for classification [7]. Convolutional neural networks is a subclass of deep neural networks, are most typically employed to evaluate visual pictures. It employs a special technique known as convolution. Convolution is a mathematical operation that takes two functions and produces a third function that describes how the forms of the first two are altered. Convolutional neural networks are made up of many layers of artificial neurons [15]. Fig. 3 shows the architecture of CNN classifier. The following provides a succinct description of the many layers employed by CNN for classification.

a) Convolutional Layer: The foundational component of the CNN is the convolution layer. Most of the computational load on the network is carried by it. In this layer, two matrices are merged to generate a dot product, which includes the kernel, a set of learnable parameters, and the restricted area of the receptive field. In comparison to a picture, the kernel is lesser in depth. This implies that if an image has three RGB channels, the kernel height and width will be spatially small but the depth will increase to take into account all three channels. The kernel travels the height and breadth of the image during the forward pass, creating an image of that receptive region. The resulting activation map, a two-dimensional representation of the image, reveals the kernel's reaction at each location in the image.

b) Pooling Layer: Following the convolutional layers, this layer is used. Building local areas that are then integrated into a single output convolutional feature map is the aim of the pooling layers. Max-pooling and average pooling are the two most used pooling operators. The maximum filter activation from various points inside a quantized window is used by the max-pooling layer to produce a lower resolution version of the convolutional layer activations.

c) ReLu: Rectified linear unit it is known as Relu. In deep neural networks or multi- layer neural networks, its activation function is nonlinear. The Relu layer removes all negative values from the filtered image and substitutes them with zero when the input is less than zero, the output is zero. This function is only activated when a certain threshold is crossed by the node input. Once the input rises above a particular threshold, the dependent variable and the input have a linear relationship. The ReLU function's primary benefit over other activation functions is that it does not simultaneously fire all of the neurons. When ReLU is used, the exponential increase of the computations needed to run the neural network is reduced. The computational cost of adding additional ReLU's rises linearly as the CNN's size scales.

d) Batch Normalization Layer: The layer of batch normalization enables the network's layers to learn more independently. The output of the earlier layers is normalized using it. In normalization, the activations scale the input layer. When training, particularly deep neural networks, using the batch normalization method, the contributions to a layer for each mini-batch are normalized. As a result, the learning process is stabilized and there is a significant reduction in the number of training epochs required to build deep neural networks.

e) Dropout Layer: The Dropout layer serves as a mask, keeping all other neurons viable but removing some neurons' contributions to the following layer. Dropout layers are essential in the training of CNNs because they prevent overfitting on the training data. The first set of training examples has an overly big impact on learning if they aren't present. This would prevent features from being learned that only appear in later samples or batches.

f) SoftMax Layer: The output layer of neural network uses the softmax function as the activation function. In multi-class, categorization problems occurs when there are more than two class labels that require class membership. SoftMax is utilized as the activation function. The softmax layer is useful in this case since it converts the scores into a normalized probability distribution that may be displayed to a user or used as input to other systems. For this reason, the final layer of the neural network is typically a softmax function.

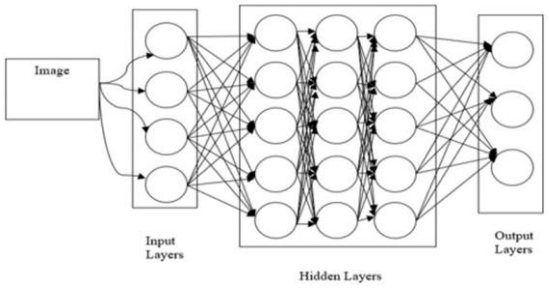

Fig. 3. CNN architecture [1]

IV. RESULTS AND DISCUSSION

In this work, python 3.8 is used for simulation. Some libraries like Pandas, NumPy are installed for training and testing in order to classify and recognize the emotions. The available data is divided in 80:20 ratio for training and testing purpose. Table I shows the count of samples in each decision and division of samples for training and testing. Also, the number of features extracted in each class.

Table. I Count of Samples

Class	No. of Samples	No. of Features/ Samples	Training		Testing	
			No. of Samples	No. of Features	No. of Samples	No. of Features
Female angry	192	13	155	2015	37	481
Female disgust	192	13	151	1963	41	533
Female fear	192	13	149	1937	43	559
Female happy	192	13	149	1937	43	559
Female neutral	288	13	234	3042	54	702
Female sad	192	13	168	2184	24	312
Female surprise	192	13	149	1937	43	559
Male angry	192	13	150	1950	43	546
Male disgust	192	13	149	1937	43	559
Male fear	192	13	155	2015	37	481
Male happy	192	13	153	1989	39	507
Male neutral	288	13	231	3003	57	741
Male sad	192	13	155	2015	37	481
Male surprise	192	13	156	2028	36	468
Total	2880		2304		576	

The results of proposed approach on different speech signals is analyzed. The performance of the model is determined by the accuracy and the loss rate.

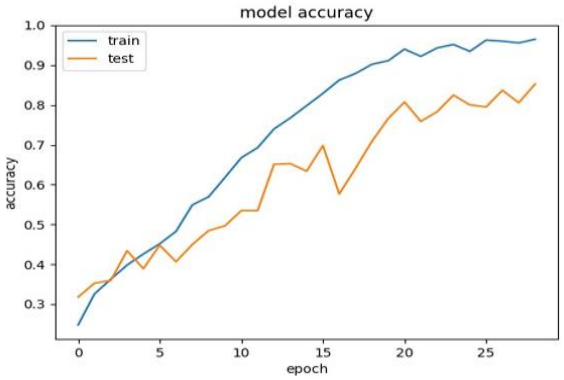

Fig. 4. Accuracy of the model

The accuracy and loss of the model are plotted against the number of epochs. Fig. 4 shows the number of epoch increase, accuracy rate is also increases.

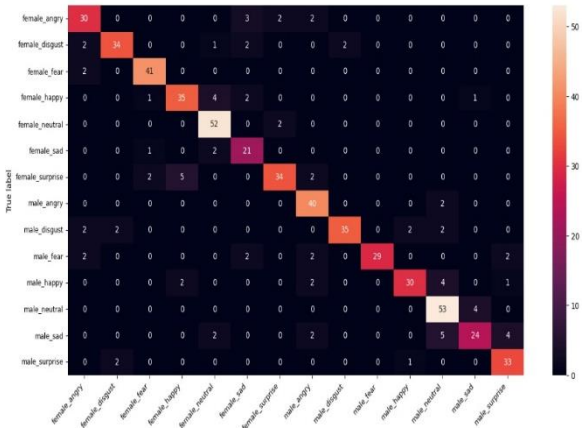

Fig. 5. Confusion matrix

The confusion matrix shows the prediction accuracy of each class against the actual class. Table II and Table III indicate different performance indicators for CNN classification model. From Table III is observed that the CNN classifier is giving high accuracy for all the emotions considered in this work. The accuracies of all classes are comparable obtaining overall classification accuracy of 97.86%. Hence it can be concluded that MFCC features are suitable for emotion recognition in using speech samples. The MFCC features are calculated using non-linear frequency scale called as mel scale using critical band filter bank which are similar to the critical bands in the human ear. Hence MFCC feature mimics the human auditory perception and works in a similar way in perceiving the emotional voice as done by human ear. MFCC also de-convolves vocal tract and vocal fold responses helping us to model each of these parameters in a better manner.

Table. II Test result

Classes	TP	FP	FN	TN
Female angry	30	8	7	461
Female disgust	34	4	7	457
Female fear	41	4	2	450
Female happy	35	7	8	456
Female neutral	52	9	2	439
Female sad	21	9	3	470
Female surprise	34	4	9	457
Male angry	40	10	2	451
Male disgust	35	2	8	456
Male fear	29	0	8	462
Male happy	30	3	9	461
Male neutral	53	13	4	438
Male sad	24	5	13	467
Male surprise	33	7	3	458

Table IV shows the comparison of the proposed methodology with already published work in the literature. It is observed that most of the work is carried out using IEMOCAP and Berlin IEMOCA EmoDB databases. The accuracy obtained for these databases using different combinations of CNN, LSTM and DBN are lower than the overall accuracy obtained using CNN model for RAVDEES database for seven distinct emotions. Dividing the speech samples further based on gender have contributed for improvement in accuracy. As the speech features vary with respect to the structure of vocal folds and vocal tract which is different for different gender, implementing gender wise classifier will rule out the gender specific variations present in the speech signal [20].

Table. III Performance Evaluation of Parameters

Classes	TPR	FPR	FNR	TNR	Precision	Recall	F1_score	Support	Accuracy (%)
Female angry	0.81	0.017	0.19	0.98	0.79	0.81	0.80	37	97.03
Female disgust	0.83	0.009	0.17	0.99	0.89	0.83	0.86	41	98.09
Female fear	0.95	0.009	0.05	0.99	0.91	0.95	0.93	43	98.95
Female happy	0.81	0.015	0.19	0.98	0.83	0.81	0.82	43	97.39
Female neutral	0.96	0.020	0.04	0.98	0.85	0.96	0.90	54	98.09
Female sad	0.88	0.018	0.13	0.98	0.70	0.78	0.78	24	97.91
Female surprise	0.79	0.009	0.21	0.98	0.89	0.79	0.84	43	97.74
Male angry	0.95	0.021	0.05	0.98	0.80	0.95	0.87	42	97.91
Male disgust	0.81	0.004	0.19	0.99	0.95	0.81	0.88	43	98.26
Male fear	0.78	0.0	0.22	1.00	1.00	0.78	0.88	37	98.61
Male happy	0.77	0.006	0.23	0.99	0.91	0.77	0.83	39	97.91
Male neutral	0.93	0.029	0.07	0.97	0.80	0.93	0.86	57	97.05
Male sad	0.65	0.010	0.35	0.99	0.83	0.65	0.73	37	96.87
Male surprise	0.92	0.015	0.08	0.98	0.82	0.92	0.87	36	98.26

Table. IV Comparison of proposed method with other research works in the domain of speech emotion recognition

S. No	Emotion Recognized	Database	Approach	Accuracy
1	Fear, Surprise, Happiness, Sadness, Neutral, Disgust and Anger. [16]	Berlin EmoDB and IEMOCAP	LSTM, DBN and CNN	An accuracy of 91.6% and 92.9% is obtained with Deep 1D and 2D CNN LSTM
2	Happiness, Neutral, Anger and Sadness. [17]	IEMOCAP database	RNN -LSTM based with 3 layers	Overall accuracy of 71.04%
3	Surprise, Happiness, Sadness, Disgust, Fear, and Anger and Neutral. [18]	IEMOCAP database	2D CNN with Phoneme data input data	Achieve accuracy of about 4% above average accuracy of existing methods
4	Anger, Neutral, Happiness, and Sadness. [19]	IEMOCAP database	RNN and CNN	An accuracy of 83.2% obtained with combined RNN-CNN
5	Neutral, happy, surprise, sad, angry, disgust, fear. (proposed method)	RAVDEES database	CNN with independent models for male and female for every emotion	Overall accuracy of 97.86%.

V. CONCLUSION AND FUTURE SCOPE

In this work, deep learning approach using CNN is used for classification of emotions. MFCC, which represents overall shape of spectral envelope, is given as input to the CNN model. As the shape of the spectral envelope changes significantly with the different emotions, MCC obtained significant classification accuracy in detecting the emotions in speech samples. The male speech samples are comparatively accurate than female speech samples. Future work can be done by using datasets of different languages. Implementing a model which works well on various languages can be used in variety of fields such as call centers, criminal investigation, medical field etc. Multiple features can be used to increase the accuracy of model.

REFERENCES

[1] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. "Speech emotion recognition using deep learning techniques: A review." IEEE Access 7, pp. 117327-117345, 2019.

[2] S Lugović, Ivan Dunđer, and Marko Horvat. "Techniques and applications of emotion recognition in speech." In 2016 39th international convention on information and communication technology, electronics and microelectronics (mipro), pp. 1278-1283. IEEE, 2016.

[3] Misbah Farooq, Fawad Hussain, Naveed Khan Baloch, Fawad Riasat Raja, Heejung Yu, and Yousaf Bin Zikria. "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network." Sensors 20, no. 21, pp. 6008, 2020.

[4] J. Umamaheswari, and A. Akila. "An enhanced human speech emotion recognition using hybrid of PRNN and KNN." In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), pp. 177-183. IEEE, 2019.

[5] Amol M. Patil, Dilip S. Patil, and Pravin S. Patil. "Iris recognition using gray level co-occurrence matrix and Hausdorff dimension." International Journal of Computer Applications 133, no. 8, pp. 29-34, 2016.

[6] Zhiyan Han, and Jian Wang. "Speech emotion recognition based on Gaussian kernel nonlinear proximal support vector machine." In 2017 Chinese Automation Congress (CAC), pp. 2513-2516. IEEE, 2017.

[7] Nam Kyun Kim, Jiwon Lee, Hun Kyu Ha, Geon Woo Lee, Jung Hyuk Lee, and Hong Kook Kim. "Speech emotion recognition based on multi-task learning using a convolutional neural network." In 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 704-707. IEEE, 2017.

[8] Mohan Ghai, Shamit Lal, Shivam Duggal, and Shrey Manik. "Emotion recognition on speech signals using machine learning." In 2017 international conference on big data analytics and computational intelligence (ICBDAC), pp. 34-39. IEEE, 2017.

[9] S. Basu, Chakraborty, J., Bag, A., & Aftabuddin, M. (2017, March). A review on emotion recognition using speech. In 2017 International conference on inventive communication and computational technologies (ICICCT) (pp. 109-114). IEEE

[10] Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby. "Deep learning techniques for speech emotion recognition, from databases to models." Sensors 21, no. 4, pp. 1249, 2021.

[11] Deepak Bharti, and Poonam Kukana. "A hybrid machine learning model for emotion recognition from speech signals." In 2020 international conference on smart electronics and communication (ICOSEC), pp. 491-496. IEEE, 2020.

[12] Che-Wei Huang, and Shrikanth Narayanan. "Characterizing types of convolution in deep convolutional recurrent neural networks for robust speech emotion recognition." arXiv preprint arXiv:1706.02901, 2017.

[13] Fatemeh Noroozi, Tomasz Sapiński, Dorota Kamińska, and Gholamreza Anbarjafari. "Vocal-based emotion recognition using random forests and decision tree." International Journal of Speech Technology 20, no. 2, pp. 239-246, 2017.

[14] M. S. Likitha, Sri Raksha R. Gupta, K. Hasitha, and A. Upendra Raju. "Speech based human emotion recognition using MFCC." In 2017 international conference on wireless communications, signal processing and networking (WiSPNET), pp. 2257-2260. IEEE, 2017.

[15] Siddique Latif, Rajib Rana, Shahzad Younis, Junaid Qadir, and Julien Epps. "Transfer learning for improving speech emotion classification accuracy." arXiv preprint arXiv:1801.06353 (2018).

[16] Jianfeng Zhao, Xia Mao, and Lijiang Chen. "Speech emotion recognition using deep 1D & 2D CNN LSTM networks." Biomedical signal processing and control, Vol. 47, 312-323, 2019.

[17] Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi. "Multi-modal emotion recognition on iemocap dataset using deep learning." arXiv preprint arXiv: 1804.05788 (2018).

[18] Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa. "Speech Emotion Recognition Using Spectrogram & Phoneme Embedding." In Interspeech, vol. 2018, pp. 3688-3692. 2018.

[19] Egor Lakomkin, Mohammad Ali Zamani, Cornelius Weber, Sven Magg, and Stefan Wermter. "On the robustness of speech emotion recognition for human-robot interaction with deep neural networks." In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 854-860. IEEE, 2018.

[20] Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. "Quantifying bias in automatic speech recognition." arXiv preprint arXiv:2103.15122, 2021.