Understanding Emotions with Deep Learning: A Model for Detecting Speech and Facial Expressions

Kondragunta Rama Krishnaiah, Professor, Department of Computer Science and Engineering, R K College of Engineering, Vijayawada - 521456, Andhra Pradesh, India, email: kondraguntark@gmail.com

Alahari Hanumant Prasad, Professor, Department of Computer Science and Engineering, R K College of Engineering, Vijayawada - 521456, Andhra Pradesh, India, email: hanuma.alahari@gmail.com

ABSTRACT

In recent years, significant progress has been made in the realm of artificial intelligence, machine learning, and human-machine interaction. One of the increasingly popular aspects is voice interaction, where people can command machines to perform specific tasks, and devices like smartphones and smart speakers are now integrated with voice assistants like Siri, Alexa, Cortana, and Google Assistant. However, despite these advancements, machines still have limitations when it comes to engaging in conversations with humans as true conversational partners. They struggle to recognize human emotions and respond appropriately, which is why emotion recognition from speech has become a cutting-edge research topic in the field of human-machine interaction. As our reliance on machines becomes more ingrained in our daily lives, there is a growing demand for a more robust man-machine communication system. Numerous researchers are currently dedicated to the field of speech emotion recognition (SER) with the aim of improving interactions between humans and machines. The ultimate goal is to develop computers capable of recognizing emotional states and reacting to them in a manner akin to how we humans do. To achieve this goal, the key lies in accurately extracting emotional features from speech and employing effective classifiers. In this project, our focus was on identifying four fundamental emotions: anger, sadness, neutral, and happiness from speech. To do so, we utilized a convolutional neural network (CNN) in conjunction with the Mel Frequency Cepstral Coefficient (MFCC) as the technique for feature extraction from speech. After conducting simulations, the results demonstrated the superiority of the proposed MFCC-CNN model when compared to existing approaches. This promising outcome brings us closer to realizing a more emotionally intelligent man-machine interaction system, paving the way for more natural and meaningful exchanges between humans and technology.

Keywords: Speech emotion, facial emotion, convolutional neural network, Mel Frequency Cepstral Coefficient, speech emotion recognition

1. INTRODUCTION

Automatic identification of emotions by facial expressions consists of three steps: face recognition, extraction and classification of features or hand movements, facial features, and voice sound that are used to convey emotions and input. Nonetheless, the latest developments of human user interfaces, which have progressed from traditional mouse and keyboard to automated speech recognition technologies to unique interfaces tailored for individuals with disabilities, do not take full account of these important interactive capabilities, sometimes contributing to less than normal experiences When machines were able to understand such emotional signals, they could provide users precise and effective support in ways that are more in line with the desires and expectations of the individual. From psychological science it is generally agreed that human emotions may be divided into six archetypal feelings: shock, terror, disgust, rage, joy and sadness. Facial expression and voice sound play a critical role in communicating certain emotions.

Emotion interpretation has arisen as an essential field of research that can provide some useful insight to a number of ends. People communicate their feelings through their words and facial gestures, consciously or implicitly. To interpret emotions may be used several different types of knowledge, such as voice, writing, and visual. Speech and facial expression have been the valuable tool for identifying feelings since ancient times, and have revealed numerous facets, including mentality. It is an enormous and difficult job to determine the feelings beneath these statements and facial expressions. Scientists from multiple disciplines are seeking to find an effective way to identify human emotions more effectively from different outlets, like voice and facial expressions, to tackle this issue.

Computer intelligence, natural language modelling systems, etc., have been used to gain greater precision in this responsiveness towards various speeches and vocal-based strategies. Analysis of the feelings may be effective in several specific contexts. One such area is cooperation with the human computers. Computers can make smarter choices and aid consumers with emotion recognition and can also aid render human-robot experiences more realistic. We would explore current emotion recognition methods, emotion modelling, emotion databases, their features, drawbacks, and some potential future directions in this study. We concentrate on evaluating work activities focused on voice and facial recognition to evaluate emotions. We studied different technical sets that were included in current methodologies and technologies. The essential accomplishments in the sector are completed and potential strategies for improved result are highlighted.

2. LITERATURE SURVEY

Research on FER has been gaining much attention over the past decades with the rapid development of artificial intelligence techniques. For FER systems, several feature-based methods have been studied. These approaches detect a facial region from an image and extract geometric or appearance features from the region. The geometric features generally include the relationship between facial components. Facial landmark points are representative examples of geometric features [2, 30, 31]. The global facial region features or different types of information on facial regions are extracted as appearance features [20, 36]. The global futures generally include principal component analysis, a local binary pattern histogram, and others. Several of the studies divided the facial region into specific local regions and extracted region specific appearance features [6, 9]. Among these local regions, the important regions are first determined, which results in an improvement in recognition accuracy. In recent decades, with the extensive development of deep-learning algorithms, the CNN and recurrent neural network (RNN) have been applied to the various fields of computer vision. Particularly, the CNN has achieved great results in various studies, such as face recognition, object recognition, and FER [10, 16, 44]. Although the deep-learning-based methods have achieved better results than conventional methods, micro-expressions, temporal variations of expressions, and other issues remain challenging [21].

Speech signals are some of the most natural media of human communication, and they have the merit of real-time simple measurement. Speech signals contain linguistic content and implicit paralinguistic information, including emotion, about speakers. In contrast to FER, most speech-emotion recognition methods extract acoustic features because end-to-end learning (i.e., one-dimensional CNNs) cannot extract effective features automatically compared to acoustic features. Therefore, combining appropriate audio features is key. Many studies have demonstrated the correlation between emotional voices and acoustic features [1, 5, 14, 18, 27, 32, 34]. However, because explicit and deterministic mapping between the emotional state and audio features does not exist, speech-based emotion recognition has a lower rate of recognition than other emotion-recognition methods, such as facial recognition. For this reason, finding the optimal feature set is a critical task in speech-emotion recognition.

Using speech signals and facial images can be helpful for accurate and natural recognition when a computer infers human emotions. To do this, the emotion information must be combined appropriately to various degrees. Most multimodal studies focus on three strategies: feature combination, decision fusion, and model concatenation. To combine multiple inputs, deep-learning technology, which is applied to various fields, can play a key role [7, 22]. To combine the models with different inputs, model concatenation is simple to use. Models inputting different types of data output each encoded tensor. The tensors of each model can be connected using the concatenate function. Yaxiong et al. converted speech signals into mel-spectrogram images for a 2D CNN to accept the image as input. In addition, they input the facial expression image into a 3D CNN. After concatenating the two networks, they employed a deep belief network for the highly nonlinear fusion of multimodal emotion features [28]. Decision fusion aims to process the category yielded by each model and leverage the specific criteria to re-distinguish. To do this, the SoftMax functions of the different types of networks are fused by calculating the dot product using weights where the summation of the weights is 1. Xusheng et al. proposed a bimodal fusion algorithm to realize speech-emotion recognition, where both facial expressions and speech information are optimally fused. They leveraged the MFCC to convert speech signals into features and combined the CNN and RNN models. They used the weighted-decision fusion method to fuse facial expressions and speech signals [40]. Jung et al. used two types of deep networks—the deep temporal appearance network and the deep temporal geometry network—to reflect not only temporal facial features but also temporal geometry features [17]. To improve the performance of their model, they presented the joint fine-tuning method integrating these two networks with different characteristics by adding the last layers of the fully connected layer of the networks after pre-training the networks. Because these methods mostly use shallow fusion, a more complete fusion model must be designed [28].

3. PROPOSED METHOD

Emotions are an essential aspect of communication between human beings. There is a very close relationship between emotions, behaviour, and thoughts in such a way that the combination of these aspects governs the way we act and the decisions we make. For this reason, over the past years, there has been a growing interest in this area of scientific research. Automatic recognition of emotions can be applied in several areas to enhance them. For example, human–computer interaction, since detecting the emotional state of a computer system's user will allow generating a more natural, productive, and intelligent interaction. Another area is human–human interaction monitoring, given its allowance to detect conflicts or unwanted situations. This project addresses the automatic emotion recognition from speech, face, and videos as well. The proposed methodology employed deep learning CNN such as the creation of corpora, the feature selection, the design of an appropriate classification scheme, and the fusion with other sources of information, such as text.

Figure 1 shows the proposed block diagram of face and speech-based emotion recognition. RACVDESS dataset is considered to implement this work, which contains both speech and face data files. Then, pre-processing operation is carried out on both datasets performed, which removed the noises from facial images and speech files. Then, MFCC features are extracted only from speech data. Then, CNN model is trained with the both speeches based MFCC features and pre-processed facial data. Finally, test face and speech data are applied and test features are compared with the pre-trained

CNN model features. Finally, the predicted emotion is obtained through this AI-CNN model from both face and speech data.

3.1 Dataset

For facial emotion detection model, we have used 28,709 images with 7 different emotions includes angry, happy, neutral, sad, disgusted, fearful, and surprised. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset is used for speech emotion detection model. The data rate, sample frequency, and format of speech audio-only files from the RAVDESS is 16bit, 48kHz, and .wav. This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically matched statements in a neutral North American accent. Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

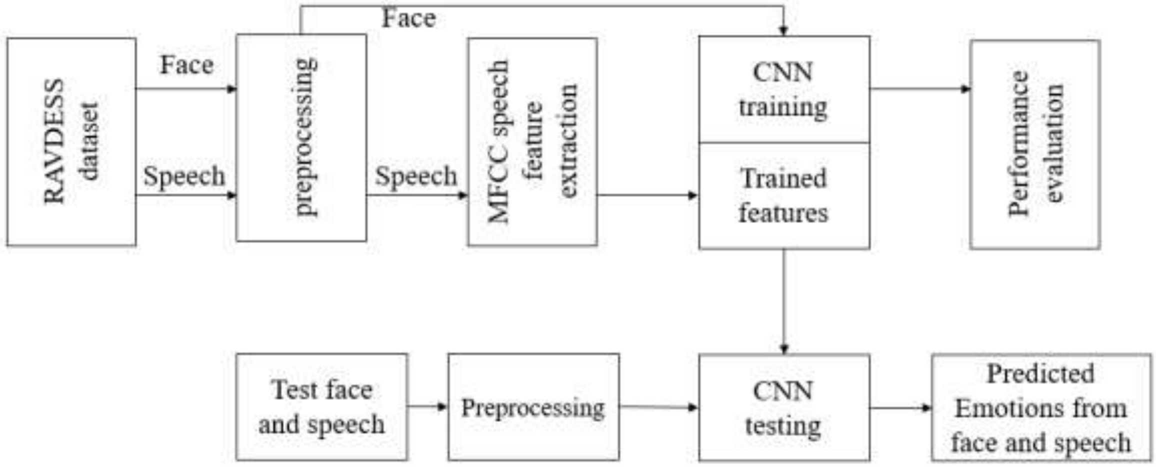

Fig. 1. Proposed block diagram

3.2 Image and Speech Pre-processing

Digital image processing is the use of computer algorithms to perform image processing on digital images. As a subfield of digital signal processing, digital image processing has many advantages over analogue image processing. It allows a much wider range of algorithms to be applied to the input data — the aim of digital image processing is to improve the image data (features) by suppressing unwanted distortions and/or enhancement of some important image features so that our AI-Computer Vision models can benefit from this improved data to work on. To train a network and make predictions on new data, our images must match the input size of the network. If we need to adjust the size of images to match the network, then we can rescale or emotion data to the required size. We can effectively increase the amount of training data by applying randomized augmentation to data. Augmentation also enables to train networks to be invariant to distortions in image data. For example, we can add randomized rotations to input images so that a network is invariant to the presence of rotation in input images. An augmented Image Datastore provides a convenient way to apply a limited set of augmentations to 2-D images for classification problems. We can store image data as a numeric array, an Image Datastore object, or a table. An Image Datastore enables to import data in batches from image collections that are too large to fit in memory. we can use an augmented image datastore or a resized 4-D array for training, prediction, and classification. We can use a resized 3-D array for prediction and classification only.

There are two ways to resize image data to match the input size of a network. Rescaling multiplies the height and width of the image by a scaling factor. If the scaling factor is not identical in the vertical and horizontal directions, then rescaling changes the spatial extents of the pixels and the aspect ratio. Cropping extracts a subregion of the image and preserves the spatial extent of each pixel. We can crop images from the center or from random positions in the image. An image is nothing more than a two-dimensional array of numbers (or pixels) ranging between 0 and 255. It is defined by the mathematical function f(x,y) where x and y are the two co-ordinates horizontally and vertically.

Resize image: In this step-in order to visualize the change, we are going to create two functions to display the images the first being a one to display one image and the second for two images. After that, we then create a function called processing that just receives the images as a parameter. Need of resize image during the pre-processing phase, some images captured by a camera and fed to our AI algorithm vary in size, therefore, we should establish a base size for all images fed into our AI algorithms.

3.2 MFCC feature extraction

Pre-emphasis is the initial stage of extraction. It is the process of boosting the energy in high frequency. It is done because the spectrum for voice segments has more energy at lower frequencies than higher frequencies. This is called spectral tilt which is caused by the nature of the glottal pulse. Boosting high-frequency energy gives more info to Acoustic Model which improves phone recognition performance. MFCC can be extracted by following method.

1) The given speech signal is divided into frames (~20 ms). The length of time between successive frames is typically 5-10ms.
2) Hamming window is used to multiply the above frames to maintain the continuity of the signal. Application of hamming window avoids Gibbs phenomenon. Hamming window is multiplied to every frame of the signal to maintain the continuity in the start and stop point of frame and to avoid hasty changes at end point. Further, hamming window is applied to each frame to collect the closest frequency component together.
3) Mel spectrum is obtained by applying Mel-scale filter bank on DFT power spectrum. Mel-filter concentrates more on the significant part of the spectrum to get data values. Mel-filter bank is a series of triangular band pass filters similar to the human auditory system. The filter bank consists of overlapping filters. Each filter output is the sum of the energy of certain frequency bands. Higher sensitivity of the human ear to lower frequencies is modeled with this procedure. The energy within the frame is also an important feature to be obtained. Compute the logarithm of the square magnitude of the output of Mel-filter bank. Human response to signal level is logarithm. Humans are less sensitive to small changes in energy at high energy than small changes at low energy. Logarithm compresses dynamic range of values.
4) Mel-scaling and smoothing (pull to right). Mel scale is approximately linear below 1 kHz and logarithmic above 1 kHz.
5) Compute the logarithm of the square magnitude of the output of Mel filter bank.
6) DCT is further stage in MFCC which converts the frequency domain signal into time domain and minimizes the redundancy in data which may neglect the smaller temporal variations in the signal. Mel-cepstrum is obtained by applying DCT on the logarithm of the mel-spectrum. DCT is used to reduce the number of feature dimensions. It reduces spectral correlation between filter bank coefficients. Low dimensionality and 17 uncorrelated features are desirable for any statistical classifier. The cepstral coefficients do not capture the energy. So, it is necessary to add energy feature. Thus twelve (12) Mel Frequency Cepstral Coefficients

plus one (1) energy coefficient are extracted. These thirteen (13) features are generally known as base features.

7) Obtain MFCC features.

The MFCC i.e., frequency transformed to the cepstral coefficients and the cepstral coefficients transformed to the MFCC by using the equation.

$$mel(f) = 2595 \times \log 10 \left(1 + \frac{f}{700}\right)$$

Where f denotes the frequency in Hz the Step followed to compute MFCC. The MFCC features are estimated by using the following equation.

$$C_n = \sum\nolimits_{n=1}^{K} (logS_k)\left[n\left(K - \frac{1}{2}\right)\frac{\pi}{K}\right] wheren = 1,2, \ldots\ldots K$$

Here, K represents the number of Mel cepstral coefficient, C0 is left out of the DCT because it represents the mean value of the input speech signal which contains no significant speech related information. For each of the frames (approx. 20 ms) of speech that has overlapped, an acoustic vector consisting of MFCC is computed. This set of coefficients represents as well as recognize the characteristics of the speech.

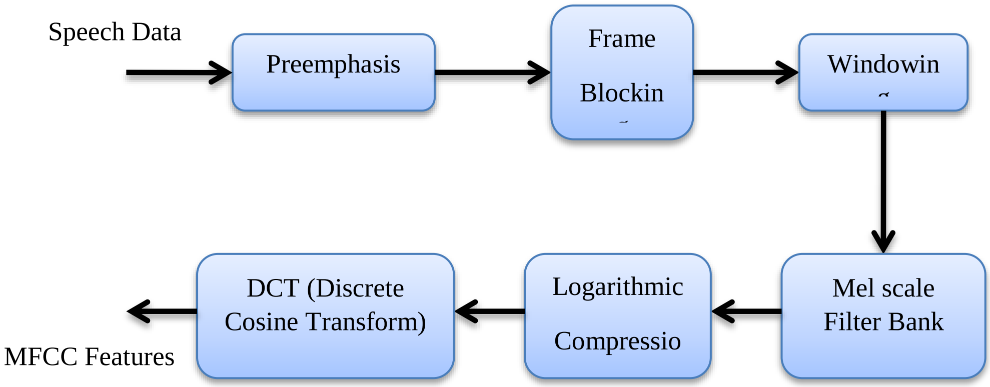

Fig. 2. MFCC operation diagram

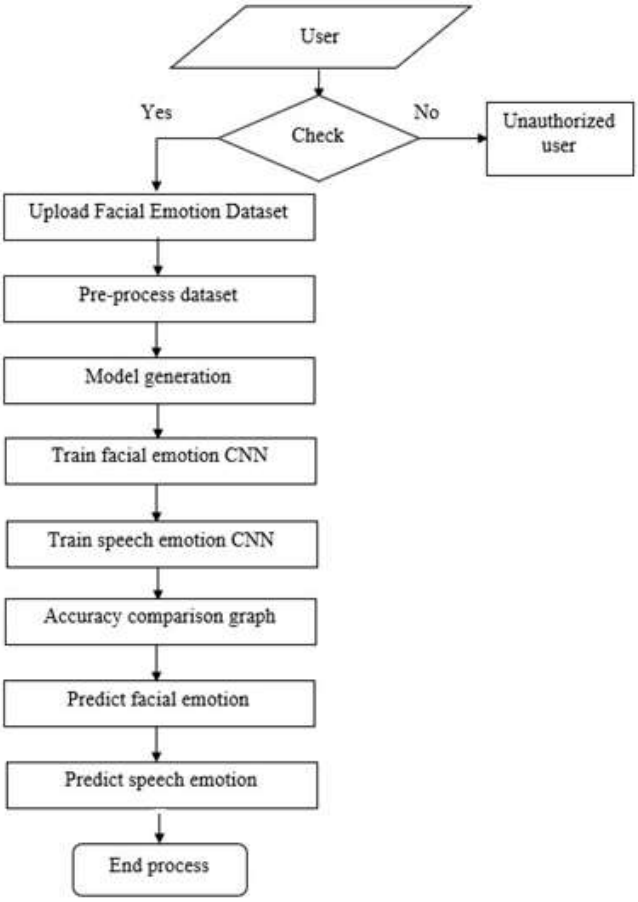

Fig. 5. Proposed data flow diagram for emotion detection model from speech, facial expression.

Figure 4 demonstrate the data flow diagram (DFD) of proposed deep CNN model. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system. It is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system. In addition, it shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output. Moreover, it may be used to represent a system at any level of abstraction, and it may be partitioned into levels that represent increasing information flow and functional detail.

4. RESULTS

Figure 6 illustrate the sample test images of emotion prediction from given facial expressions, where it includes all the emotions such as sad, angry, neutral, disgusted, surprised, and fearful. Figure 7 discloses the obtained prediction accuracy and loss performance using proposed deep CNN from facial expression, speech, and videos. From both the figures, it is observed that proposed deep CNN obtained superior performance for emotion prediction from videos as compared to both facial expression and speech inputs.

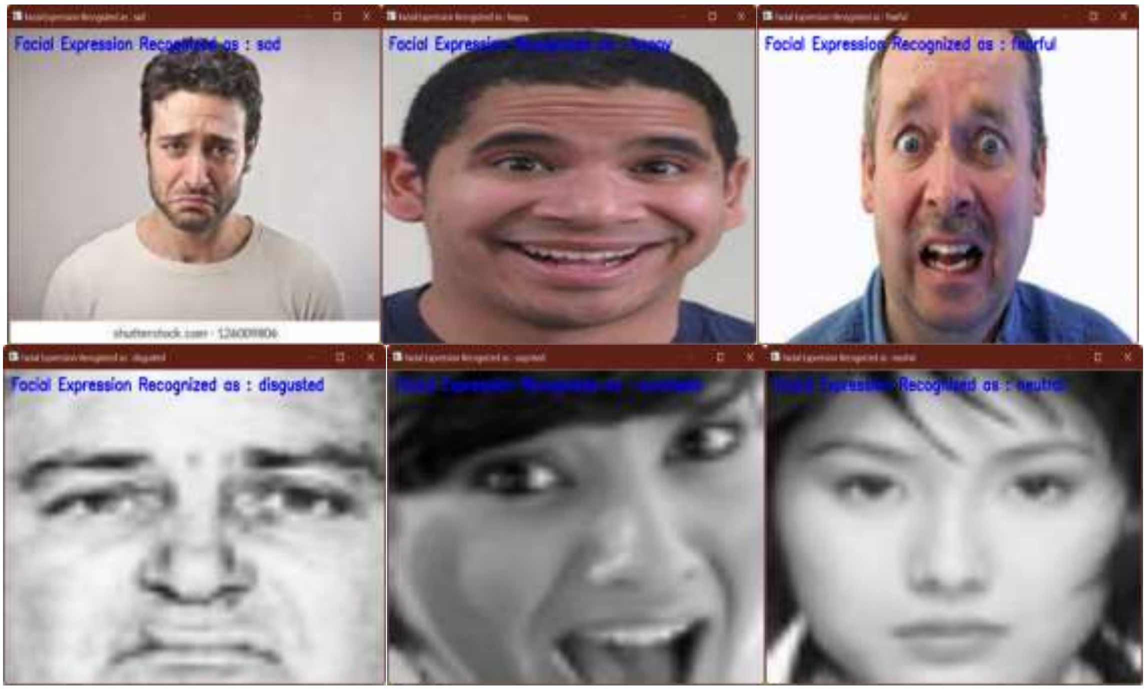
Facial Expression Recognized as : sad
Facial Expression Recognized as : happy
Facial Expression Recognized as : fearful
Facial Expression Recognized as : disgusted
Facial Expression Recognized as : surprised
Facial Expression Recognized as : neutral

Fig. 6. Sample test images of emotion prediction.

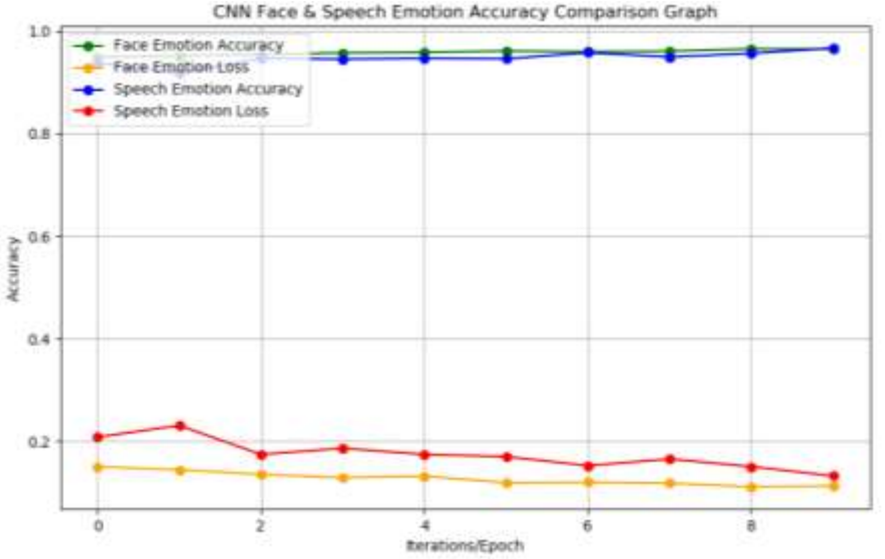

Fig. 7. Accuracy and loss comparison of proposed CNN with speech and facial expression.

5. CONCLUSION

Emotion interpretation has arisen as an essential field of research that can provide some useful insight to a number of ends. People communicate their feelings through their words and facial gestures, consciously or implicitly. To interpret emotions may be used several different types of knowledge, such as voice, writing, and visual. Therefore, this work proposed a deep CNN model for emotion prediction from speech, and facial expression with enhanced prediction accuracy and reduced loss. In addition, the speech CNN model utilized MFCC as feature extraction from given speech samples.

REFERENCES

[1] Bjorn S, Stefan S, Anton B, Alessandro V, Klaus S, Fabien R, Mohamed C, Felix W, Florian E, Erik M, Marcello M, Hugues S, Anna P, Fabio V, Samuel K (2013) Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism

[2] Deepak G, Joonwhoan L (2013) Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines. Sensors 13:7714–7734.

[3] Domínguez-Jiménez JA, Campo-Landines KC, Martínez-Santos J, Delahoz EJ, Contreras-Ortiz S (2020) A machine learning model for emotion recognition from physiological signals. Biomed Signal Proces 55:101646

[4] El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recogn 44:572–587.

[5] Eyben F, Scherer KR, Schuller BW et al (2016) The Geneva minimalistic acoustic parameter set (geMAPS) for voice research and affective computing. IEEE Trans Affect Comput 7:190–202.

[6] Ghimire D, Jeong S, Lee J, Park SH (2017) Facial expression recognition based on local region specific features and support vector machines. Multimed Tools Appl 76:7803–7821.

[7] Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press. https://www.deeplearningbook.org. Accessed 1 Mar 2020

[8] Hamm J, Kohler CG, Gur RC, Verma R (2011) Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. J Neurosci Methods 200:237–256

[9] Happy SL, George A, Routray A (2012) A real time facial expression classification system using local binary patterns. In Proc 4th Int Conf Intell Human Comput Interact 27–29:1–5

[10] Hasani B, Mahoor MH (2017) Facial expression recognition using enhanced deep 3D convolutional neural networks. IEEE Conf Comput Vision Pattern Recognit Workshops (CVPRW).

[11] He J, Li D, Bo S, Yu L (2019) Facial action unit detection with multilayer fused multi-task and multi-label deep learning network. KSII Trans Internet Inf Syst 7:5546–5559.

[12] Hossain MS, Muhammad G (2019) Emotion recognition using deep learning approach from audio–visual emotional big data. Inf Fusion 49:69–78.

[13] Hutto CJ, Eric G (2014) VADER: A parsimonious rule-based model for sentiment analysis of social media text. AAAI Publications, Eighth Int AAAI Conf Weblogs Soc Media

[14] Iliou T, Anagnostopoulos C-N (2009) Statistical evaluation of speech features for emotion recognition. In: Digital telecommunications ICDT'09 4th Int Conf IEEE 121–126

[15] Jia X, Li W, Wang Y, Hong S, Su X (2020) An action unit co-occurrence constraint 3DCNN based action unit recognition approach. KSII Trans Internet Inf Syst 14:924–942.

[16] Joseph R, Santosh D, Ross G, Ali F (2015) You Only Look Once: Unified, Real-Time Object Detection arXiv preprint arXiv:1506.02640

[17] Jung H, Lee S, Yim J, Park S, Kim J (2015) Joint fine-tuning in deep neural networks for facial expression recognition. 2015 IEEE Int Conf Comput Vision (ICCV).

[18] Kao YH, Lee LS (2006) Feature analysis for emotion recognition from Mandarin speech considering the special characteristics of Chinese language. In: InterSpeech

[19] Kaulard K, Cunningham DW, Bülthoff HH, Wallraven C (2012) The MPI facial expression database—A validated database of emotional and conversational facial expressions. PLoS One 7:e32321.

[20] Khan RA, Meyer A, Konik H, Bouakaz S (2013) Framework for reliable, real-time facial expression recognition for low resolution images. Pattern Recogn Lett 34:1159–1168.

[21] Ko BC (2018) A brief review of facial emotion recognition based on visual information. Sensors 18.

[22] LeCun Y, Bengio Y, Hinton G (2015) Deep learning, Nature 521.

[23] Lee C, Lui S, So C (2014) Visualization of time-varying joint development of pitch and dynamics for speech emotion recognition. J Acoust Soc Am 135:2422.

[24] Li S, Deng W (2020) Deep facial expression recognition: A survey. IEEE Trans Affective Comp (Early Access).

[25] Liu M, Li S, Shan S, Wang R, and Chen X (2014) Deeply learning deformable facial action parts model for dynamic expression analysis. 2014 Asian Conference on Computer Vision (ACCV) 143–157.

[26] Lotfian R, Busso C (2019) Curriculum learning for speech emotion recognition from crowdsourced labels. IEEE/ACM Trans Audio, Speech Lang Processing 4.

[27] Luengo I, Navas E, Hernáez I, Sánchez J (2005) Automatic emotion recognition using prosodic parameters. In: Interspeech, 493–496

[28] Ma Y, Hao Y, Chen M, Chen J, Lu P, Košir A (2019) Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. Inf Fusion 46:184–192.

[29] Mehrabian A (1968) Communication without words. Psychol Today 2:53–56

[30] Mira J, ByoungChul K, JaeYeal N (2016) Facial landmark detection based on an ensemble of local weighted regressors during real driving situation. Int Conf Pattern Recognit 1–6.

[31] Mira J, ByoungChul K, Sooyeong K, JaeYeal N (2018) Driver facial landmark detection in real driving situations. IEEE Trans Circuits Syst Video Technol 28:2753–2767.

[32] Rao KS, Koolagudi SG, Vempada RR (2013) Emotion recognition from speech using global and local prosodic features. Int J Speech Technol 16(2):143–160

[33] Scherer KR (2003) Vocal communication of emotion: A review of research paradigms. Speech Comm 40:227–256.

[34] Schuller B, Batliner A, Steidl S, Seppi D (2011) Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Comm 53(9–10):1062–1087.

[35] Shaqr FA, Duwairi R, Al-Ayyou M (2019) Recognizing emotion from speech based on age and gender using hierarchical models. Procedia Comput. Sci. 151:37–44.

[36] Siddiqi MH, Ali R, Khan AM, Park YT, Lee S (2015) Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields. IEEE Trans Image Proc 24:1386–1398.

[37] Song P, Zheng W (2018) Feature selection-based transfer subspace learning for speech emotion recognition. IEEE Trans. Affective Comput. (Early Access)

[38] Sun N, Qi L, Huan R, Liu J, Han G (2019) Deep spatial-temporal feature fusion for facial expression recognition in static images. Pattern Recogn Lett 119:49–61.

[39] Swain M, Routray A, Kabisatpathy P (2018) Databases, features and classifiers for speech emotion recognition: A review. Int J Speech Technol 21:93–120.

[40] Wang X, Chen X, Cao C (2020) Human emotion recognition by optimally fusing facial expression and speech feature. Signal Process Image Commun.

[41] Wu CH, Yeh JF, Chuang ZJ (2009) Emotion perception and recognition from speech, Affective Inf Processing 93–110.

[42] Xiong X and Fernando DlT (2013) Supervised descent method and its applications to face alignment. 2013 IEEE Conf Comput Vision and Pattern Recognit (CVPR).

[43] Zamil AAA, Hasan S, Baki SJ, Adam J, Zaman I (2019) Emotion detection from speech signals using voting mechanism on classified frames. 2019 Int Conf Robotics, Electr Signal Processing Technol (ICREST).

[44] Zhang H, Huang B, Tian G (2020) Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture. Pattern Recogn Lett 131:128–134.

[45] Zhang S, Zhang S, Huang T, Gao W (2008) Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. IEEE Trans. Multimed. 20:1576–1590.

[46] Zhang T, Zheng W, Cui Z, Zong Y, Yan J, Yan K (2016) A deep neural network-driven feature learning method for multi-view facial expression recognition. IEEE Trans. Multimed. 18:2528–2536.

[47] Zhao J, Mao X, Chen L (2019) Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomed Signal Processing Control 47:312–323.