Advances in Nonlinear Variational Inequalities
ISSN: 1092-910X
Vol 28 No. 2 (2025)

CNN-Based Voice Recognition by Self-Governing Robots to Improve Computer-Human Communication

Ben Sujin. B[1], M. Sakthivel[2], T.P.S. Kumar Kusumanchi[3], Dr.G.Naga Jyothi[4], Divya Muralitharan[5]

[1]Computer Engineering Department, University of Technology and Applied Sciences, Nizwa, Sultanate of Oman, bennet@utas.edu.om

[2]Department of Artificial Intelligence and Data Science, Sri Shanmugha College of Engineering and Technology, Sankari, Salem-637304, India, sakthisalem@gmail.com

[3]Dept of Internet of Things, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, 522302, Andhra Pradesh, India, satishkumar8421@gmail.com

[4]Asst. Professor, Madanapalle Institute of Technology and Science, Angallu, Andhra Pradesh, India, nagajyothisai221@gmail.com

[5]Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India, divyamuralitharan@gmail.com

Article History:

Received: *22-07-2024*

Revised: *03-09-2024*

Accepted: *13-09-2024*

Abstract

Robots must have the ability to recognize human emotions to engage with individuals and to plan their movements autonomously. Nonverbal signals encompassing pitch, loudness, spectrum, and speech speed are successful methods for transmitting emotions to most individuals. Provided this situation, a machine could have the capability of qualified figure out emotions by deploying the traits of spoken communication, and these potentially propagate vital information concerning the speaker's emotional state. More precisely, a combination of numerous facial action units can be employed to describe a human's emotion. In this paper, we propose a deep Convolutional Neural Network-based system that can identify sentiments in real-time and with a high accuracy rate. This investigation establishes an entirely novel speech-emotion detection system founded on Convolutional Neural Networks (CNNs). With the support of a top-tier GPU, a model is developed and nourished raw speech from a specific set for training, classification, and testing purposes. We further analyze the speech data and incorporate the information from the visual and audio sources to enhance the recognition system's accuracy. The benefits of the proposed method for emotion identification and the implications of blending visual and aural suggestions are made clear by the experimental results. Convolutional neural networks (CNNs) are taught on grayscale images from the softmax dataset in the present work. To acquire the best accuracy, we experimented with different baselines and max pooling layers, finally acquiring 89.98% accuracy. Dropout is one approach we have used to ward off overfitting.

Keywords: Deep Convolutional Neural Network; Human-computer interaction; Speech Emotion recognition; Autonomous robot.

1. Introduction

Conversations between individuals are saturated with emotions. They can alter the purpose of affirmations in addition to conveying the speaker's emotional state. It is immensely simple for humans to make use of paralinguistic, verbal, and non-verbal signals to recognize emotions in other

living forms along with in other humans courtesy of generations of evolution, but it is not as simple to infer the same from machines. Speech emotion recognition (SER) systems engage with speech paralinguistic patterns to derive the speaker's emotional state. An extensive spectrum of fields, ranging from human-robot contact, voice assistants, dialogue systems, medical science for the diagnosis of depression and stress, call centers for the assessment of staff performance and customer satisfaction, human–robot interaction, etc., might employ SER. An artificial neural network with feed-forward deep learning is called a CNN.

Fully connected layers, pooling layers, and convolutional layers contribute to a basic framework. Filters are convolved, or moved, over the input, in the convolutional layer. To improve the resolution of the feature map generated via the convolutional layer, pooling layers are implemented. Furthermore, CNNs have the capability of having numerous tightly connected layers, where virtually every layer's neurons have a connection to every layer's neuron [1]. Emotion classification has witnessed countless strategies and techniques in recent years, but designing an automated system to communicate this task is tricky. To accomplish the assignment of emotion recognition in this project, we have suggested Shallow and Deep Convolutional Neural Networks.

Image is supplied as input to our framework which subsequently predicts the passion. The states of mind are categorized into seven classes such as frustration, joy, panic, sorrow, disgrace, amazement, and neutral [2]. Automatic voice recognition pertains to a computer's capability to discern, "receive, and interpret" speech and decode it into text or understandable procedure. The skill of a CPU to listen to dialogue and spread an achievement by reacting to a human’s commands is termed as automatic speech acknowledgment. Three steps play a role in processing a phoneme: perceiving the speaker, understanding the words which are spoken, and analyzing the emotion. The words can be observed in writing or by technological devices. They are viewed as accurate knowledge, very similar to our professional experiences. The three distinct groups for spoken word extraction rely on the specific type of signal and its length, illustrated in Figure 1.1.

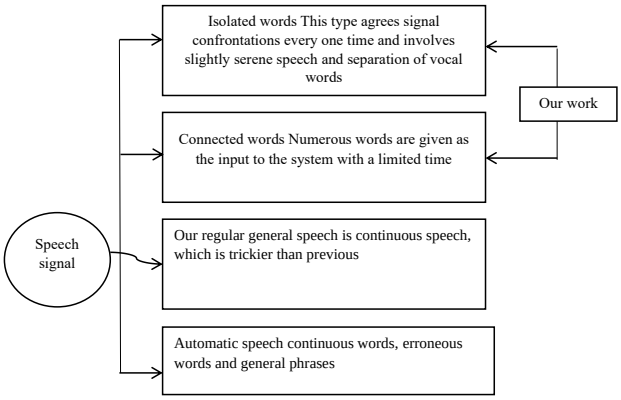

Fig. 1.1. Speech varieties

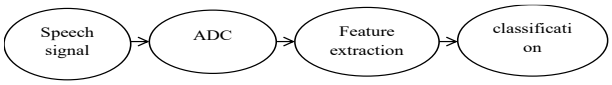

Fig. 1.2. The comprehensive way to speech recognition completion.

When creating a robotic application for home security systems or an automotive control circuit, sound recognition is crucial. In most situations, sound waves are described as longitudinal waves that go through an adiabatic compression and decompression phase. As they travel, longitudinal waves vibrate in the same direction. Spectrophotograms, which portray frequency on the vertical axis and time on the horizontal axis to represent signal energy, are two-dimensional patterns utilized with

sound processing. The consonant, which gives rise to various noises, usually happens by a flow of air in the vocal tract. Figure 1.2 presents a clarification of the general speech recognition completion approach. ADC (analog to digital converter) (spectral shaping) is the standard approach to accomplishing speech recognition theory. This entails initially choosing the sound wave, converting it into a digital form, preemphasis filtering, and feature extraction. As a targeted stage, the features obtained will be processed and sent to groupings. A separation and concatenation operation is utilized for converting obtained attributes into signal parameters in some techniques. This manner of operation is usually referred to as parameter transformation. A small component of statistical modeling necessitates parameter conversion in signal observation vectors. In security and access control systems, sound recognition is serious. A reminiscent sinusoidal movement of loud tones resonating quickly and at greater intensities than low tones is termed a sound wave. The microphone updates the resonant force of sound into electrical energy. An initial sense of the energy corresponding to the signal amplitude can be gained through the universal shape of the sound wave. Given that the nature of a sound wave is that its frequencies undergo shifts and are adherent to adjacent ones, the FFT examines the frequency components that encompass the sound wave and depicts these components via a spectrum diagram [3].

The format that follows is the overall subject matter sequence for the article: Threat verification using time series and convolutional neural networks is tackled in Section 3; experimental outcomes of time series analysis for cyber security are examined in Section 4; and a conclusion will be offered in Section 5.

2. Related Works

Deshmukh, R. S. et. al [4] The demand for computers to be able to recognize and respond to human words and in addition to behavioral signs expressing emotions and mental states has developed as an impact of the increasing prevalence of computers and computer interfaces toward everyday life. Furthermore, the facial expressions of a photo have statistical importance. In particular, the expressions that are not associated with the basic set of emotions merely connect to the physical and mental functioning of the mind. In this empirical investigation, facial expressions from live webcam pictures were taken for recognition of mental states in human snapshots. The proposed approach does not affect parameters that incorporate birthmarks, gender, age, culture, beard, or background. The proposed strategy was designed to be highly successful and produced to offer music therapy to individuals who go through stress at work.

Bai, X. et. al [5] This guest editorial summarizes the thirty recognized papers for this special issue and contains an in-depth evaluation of the representative works and the latest developments in transparent deep learning for trustworthy and efficient pattern recognition. The accepted papers are of outstanding quality, sacrificing the most recent progress in developing compact and reliable network architectures for specific pattern recognition problems, boosting the comprehensible nature of deep learning approaches, building creative ideas for adversarial attacks, and exploring training stability. These accepted stuff deliver new ideas that may encourage additional studies in this particular area. The growing recognition of AI's ethics, prejudgment, and trust is illustrated by the recent interest in explainable AI, mainly among governments under the light of the European General

Data Protection law (GDPR) law. We are convinced that this unique issue and the guest article will contribute to advancing the promising field's forward progress.

Choudhary, R. R. et. al [6] We offered methods for sorting emotions based on deep neural networks and machine learning by deploying the RAVDESS and TESS datasets. The models had been taught to sense separate feelings, leading to an overall accuracy of 0.97. With a rating of 0.90, the sorrowful class performed the best, while the shocked and disgusted class achieved the least (0.77). MFCC attributes, or the spectrum of an a-spectrum, were removed from the training audio samples. We found out that the random forest achieved an accuracy of 0.96 and the SVM had an accuracy of 0.86. Employing the deep learning algorithms LSTM, GRU, and CNN, we acquired accuracy readings of 0.92, 0.93, and 0.971, in that sequence. Despite the intriguing accomplishments, we can claim that deep neural networks serve as an excellent foundation to rectify the problem. The efficiency, exactness, and functionality of the suggested project could all be advanced in the future. Two additional elements that might have been incorporated into the model's capabilities are mood swings and discouragement. These strategies can be employed by therapists to observe their patients' emotions. It's tricky to incorporate an evocative computer with a sarcasm detection system.

Vryzas, N. et. al [7] We suggested a continuous SER system centered around CNN architecture in the current jobs. AESDD, a widely available original database of Greek-language emotional speech, was implemented to train the model. The opaque SER task was assessed for human-level accuracy by employing informal assessment experiments. In terms of accuracy, the CNN system performs more than the baseline SVM models leveraging individually generated audio features. Despite data augmentation having been anticipated to enhance flexibility and generalization, it exhibited little effect on classification precision in the validation tests. In addition to boosting performance, the advocated topology's unsupervised feature extraction portion supports the construction of real-time systems. When an extensive data set is available for initial training and a more narrow domain-specific dataset is utilized for fine-tuning, this can be quite efficient. Many techniques have been put forth for audio recognition that combines distinct audio event classification tasks.

Lakomkin, E. et. al [8] Here, we investigated two neural speech emotion detection models and revealed that, when trained exclusively using clean, in-domain data, and tested on the iCub robot, they perform significantly worse. We observed that, even in the lack of genuine robot ego noise or general situations data augmentation drastically reduced this accuracy loss and strengthened the model's overall robustness. By incorporating noise during training, we witnessed significant boosts in performance on the IEMOCAP-iCub data. Thus, we can deduce that in succession to obtain models prepared to enable robot deployment, training data augmentations needed to be implemented. Also, as it becomes difficult to judge valence without it, we are looking at the likelihood of reinforcing the input representation with spoken text information in areas of high noise.

Ruiz-Garcia, A. et. al [9] We have proven an ensemble of emotion recognition models in this work which were taught and tested against the KDEF dataset. On the KDEF dataset, our best model—a CNN and SVM combination—produced results that were almost on par with larger models and exceeded the state-of-the-art performance rate. It needs to be emphasized that neither of the CK+ images was used to train the CNN segment of the model when it was examined on the CK+ dataset; this mixture of CNN+SVM architecture falls 3.73% short of the state-of-the-art but issued

substantially better results than a larger model that was offered. Nevertheless, our hybrid model has smaller model parameters, converges more rapidly, and consumes less data to train. Apart from that, this work has established the benefits of SVM over MLP for feature classification, at least with the goal of emotion documentation, and the efficiency of CNN over Gabor filters in terms of feature extraction. In juxtaposition with similar approaches that also leverage a CNN, this hybrid architecture is intriguing since they have its synthesized configuration demands not as many hyperparameters, is relatively faster to train against the state-of-the-art, and may learn from smaller data sets.

Khan, A. et. al [10] In an overview, there are numerous examples of key concepts of autonomous entities, as well as novel theories and design architectures that derive from an assortment of these theories. In some manner or a different one, they all connect perception or detection with action and environment interaction. We are putting out the Mimicking Human Sensing Theory when illuminated by these present ideas. Following this theory, an agent must demonstrate autonomy in every one of the five senses—sight, opinions, touch, hearing, and smell—in stability to reach full self-reliance. Integrating a broad spectrum of sensors may facilitate these numerous sensing channels, which are going to help in fulfilling the tasks designated. Each of the previously mentioned senses requires to be drawn meticulously and carefully. Our deep learning project focuses on speech. NPR has been chosen as a believable data source to make sure we can acquire information from pioneers and experts. In this project, we invented an original technique in which a robot monitors a debate and can interrupt a speaker if it projects the answer to a question posed by the patient or physician. The conversation is listened to by the robot, which then breaks down the data into two primary elements: the conversation's focus and the comments to the questions on that topic.

Tai, L. et. al [11] In this study, we completed practical tests in ubiquitous indoor spaces and claimed a human-like indoor exploration algorithm based on a single deep-network structure. Experiments demonstrate that our system could do a good job averting hazards. Automated and human decision-making were assessed and the findings were extremely similar. Although numerous limitations still occur. For instance, robot applications may not be adequately delivered by the offline training method and a discrete classification may not be exact enough for a continuous state range of the decisions. In the following stages, we will progressively expand the target space from discrete space to continuous space and further en-couple online learning algorithms using libcnn.

3. Methods and Materials

In this research investigation, we provide a three-part deep Convolutional Neural Network (CNN)-based architecture. The network is broken down into three distinct parts: the first derives features from image sequences, the second pulls aspects from audio supplies, and the last segment, which is the third, conducts emotion recognition. The convolution operation is an essential element of the recommended network architecture and is defined as follows for the visual (2-D) and senses (1-D) signals:

$$(g * i)[j] = \sum_{l=-U}^{U} i[l].\, g[j-l] \quad (1)$$

$$(g * i)[j,k] = \sum_{l=-U}^{U} \sum_{n=-U}^{U} i[l,n]\,.\, g[j-l,k-n] \quad (2)$$

where g is the current 1-D or 2-D signal and $i[l]$ and $i[l,n]$ are the 1-D and 2-D kernels, correspondingly, whose parameters were acquired during the training phase.

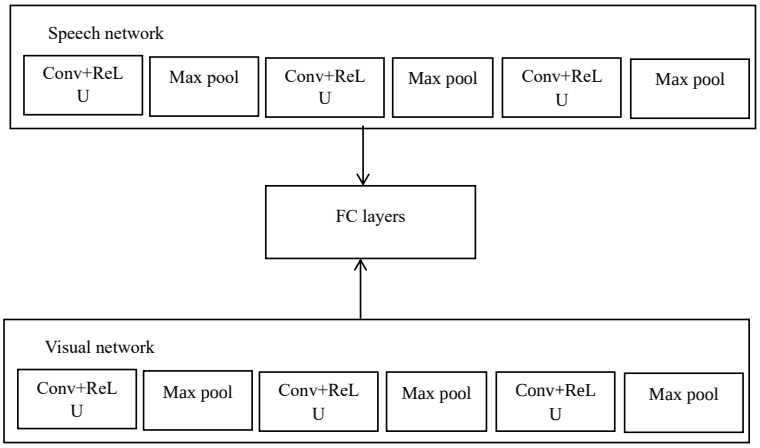

Fig. 3.1. Projected structure for emotion acknowledgment

The lower branch of the sketch exhibited in Figure 3.1 indicates the initial component of the network. It intends to process an assortment of images and synthesize abstract features that represent the information provided in every frame. The input measurements are C × O × X × I, where O is the video sequence's length, X is its width, and I is a comparable height. Each movie frame is researched personally. The primary goal of the network's second group, which is similar to CNN-based architecture, is to process audio signal spectrograms and extract pertinent details from them. At this juncture, two specifics should be emphasized. To begin with, there aren't numerous convolutional layers. Second, in comparison with the first zone of the network that conducts image analysis, the last layer has fewer dollars neurons. This is addressed by the fact that bodily information—such as the arrangement of a person's mouth, eyes, eyebrows, and cheekbones—allows more information than hearing inputs. To be more accurate, implicit signals prove essential in establishing a person’s emotional state. By way of example, a person's happiness can be immediately observed from their facial expression. Despite this, anger perception necessitates vocal expression. A classifier constructed of two fully-connected (FC) layers serves as the third portion of the network represented in Figure 3.2. The abstract attributes are concatenated after the inference through the previously discussed network elements, and the corresponding feature vector is pushed in the last phase of the network, which results in the probability spread of the emotion in the researched video. As already stated, because the visual data turns out to be more essential, the feature vector lengths for segments and sound differ by a proportion of 4:1.

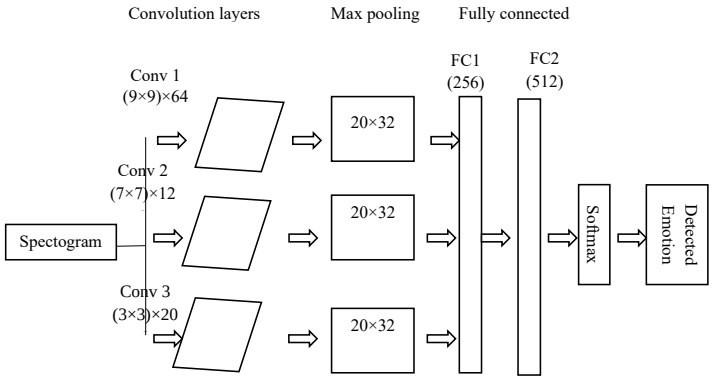

Fig. 3.2. CNN architecture for conversation emotion recognition is portrayed

The end outcome is a tensor of M components $z_{y,1}, z_{y,2}, \dots, z_{y,M}$, that symbolizes an individual's emotional state (M being the entire number of emotional states explored). We check out leveraging a SoftMax function to compute the probability distribution for the emotional states. This function makes use of the formula to alter the graph of 6 scores $z_{y,1}, z_{y,2}, \dots, z_{y,M}$ into a normalized vector of likelihood $q_{y,1}, q_{y,2}, \dots, q_{y,M}$.

$$q_{y,d} = \frac{e^{z_{y,d}}}{\sum_{d'=1}^{M} e^{z_{y,d'}}} \quad (3)$$

For $d \in \{1, \dots, M\}$

The CrossEntropy loss function, which is defined in the following manner for a measurement x, is the one utilized for training.

$$\mathcal{M}(y) = -\sum_{d=1}^{M} \varepsilon_{y,d} \log\left(q_{y,d}\right) \quad (4)$$

where $q_{y,d}$ is the estimated probability that observation y belongs to class c and $\varepsilon_{y,d}$ is the binary confirmation (0 or 1) denoting the likelihood that class label d corresponds to the reality for observation y [12].

4. Implementation and Results

Six classes of audio-acquired features correspond to the six basic sensations in the training dataset. The feature was constructed of 400 lines, each with 12 MFCC coefficients, that were chosen from a list of 200 wav files that remained voice recorded at 191 kbps and comprised 30 Romanian speaker recordings that extended 5 seconds. Using the PRAAT script deployment, voices are analyzed utilizing a 26 ms Hamming screen and an 11 ms frame rate. The recordings demonstrate approximately equal proportions of the six emotions, as stated by human operators: happiness, surprise, disgust, frustration, fear, and mourning. Below Table 1 indicates the number of files in the train/ evaluation over each emotion which is clearly stated.

Table 1. Number of files in the train/evaluation over each emotion

Joy	Fear	Frustration	Worry	Disgust	Surprise
36/6	32/6	33/6	34/6	36/6	35/6

Later the contrary, the CNN model was evaluated employing a test set of 30 voice samples, and it delivered a median precision of 71.37%, which is identical to the final results of speech recognition. The consequences of employing the audio file database to train the CNN are outlined in Table 2 and Table 3. The experimental output of the CNN models according to the varying emotions is illustrated in Figure 4.1. Whereas in Figure 4.2 illustrates recognition rate using different models according to different emotions.

Table 2. Our CNN model's experimental outputs

Joy	Fear	Frustration	Worry	Disgust	Surprise
72	76	69	75	68	70

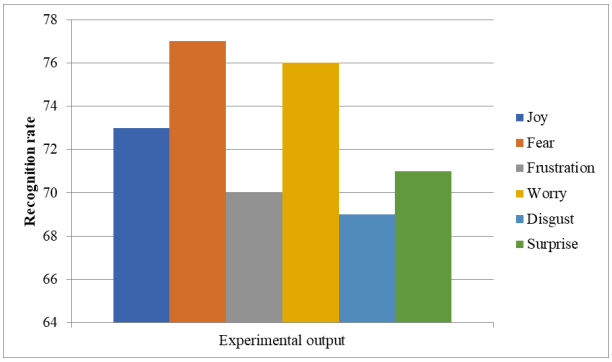

Fig. 4.1. CNN model’s experimental output according to the various emotions

Table 3. Recognition rate from scientific journals

Algorithm	Happy	Anger	Sad
Linear Discriminant Analysis (LDA)	50	69	73
Regularized Discriminant Analysis (RDA)	74	84	98
Support Vector Machines (SVM)	71	75	94
K Nearest Neighbor (KNN)	56	94	78
Convolutional Neural Networks (CNN)	72	69	75

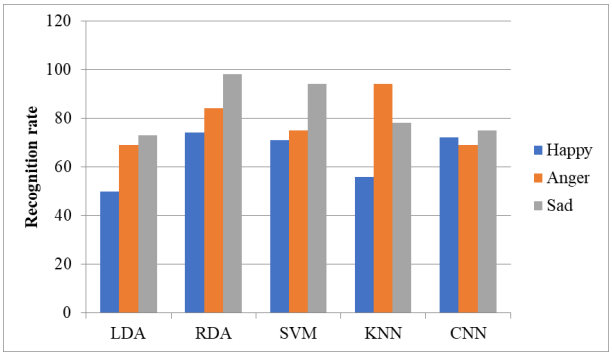

Fig. 4.2. Representation of recognition rate using different models according to different emotions

We applied ad hoc experimental planning to collect audio files for a specific feeling (delighted was our decision), for the reason to triangulate the research's results – Figure 4.2. The trial's participants recorded their voice reading a text, then proceeded to perform a short but calming activity, focused on a joyful melodies, and—optionally—watched scenes from romantic comedies or breathtaking

natural scenes on a Sony HMZ-T3 Personal 3D Viewer. Each individual then recorded their speech again. For this experiment, an unusual software module was built, and the CNN was educated by employing the files concerning the second stage. The arbitrary nature of the emotion-voice tandem whenever a human operator merely perceives the voice and has minimal additional details regarding the speaker's environment or personality is one of the troubles with emotion recognition.

The methods listed below could potentially be implemented to boost the precision of emotion detection:

1. Employ lexical analysis to check the results of linguistic parameter sorting;

2. Mix this method with further visual processing of initiatives and behaviors, or other biological qualities;

3. To utilize an extensive database to train the model and an accumulation of detailed sound clips that are expressive and analogous to each sensation to calibrate it;

4. Including an ensemble of uninfluenced recordings from an aesthetic approach.

The focus of additional investigation is to further improve the model's usefulness. The upcoming studies on ethnic and linguistic exceptions in speech parameters-based emotion identification could prove quite fascinating [13].

Matlab software was used to fabricate spectrograms from the SAVEE database. Four emotions have been put into attention. Sixty visuals were picked from the archive for each emotion. An aggregate of about 240 spectrograms were collected. Thirty percent of the data was employed for testing and 70% of the data was employed for the training phase. The training technique was carried out across three different epochs: 500, 1200, and 1500.

8 GB of RAM plus a single i5-8250 CPU were made use of for the training. We accomplished two evaluations. In the first experiment, spectrograms were implemented to train the CNN model, and the accuracy of the calculations was examined.

4.1 SER Investigation using the Suggested CNN Structure

Table 4 illustrates the numerical confusion matrix for the recommended CNN architecture with 500 training epochs. The numbers that are presented diagonally correspond to the percentage for every emotion class that was correctly identified; the remainder of the figures stands for the percentage of each sentiment class that was mistakenly identified. The prediction performance for sorrowful and neutral adults was evidently above 50%, but for angry and joyful citizens, it was below 50%. Despite this, the model's entire precision maintains at 65.5%.

Table 4. SER System Performance with 500 Epochs Using CNN Architecture

Predicted class					
Actual class		**Anger**	**Sad**	**Neutral**	**Happy**
	Anger	44.4	13.7	34.8	11.5
	Sad	10.7	79.4	1	13.2
	Neutral	4.7	0.4	94.4	3.9
	Happy	26.8	0	28	48.2

Table 5. SER System Performance with 1200 Epochs Using CNN Architecture

Predicted class					
Actual class		**Anger**	**Sad**	**Neutral**	**Happy**
	Anger	65.9	14.3	16.9	7.3
	Sad	6.9	84.4	1	11.8
	Neutral	3.9	2.3	94.4	3.8
	Happy	0	33.2	0	68.8

Table 6. SER System Performance with 1500 Epochs Using CNN Architecture

Predicted class					
Actual class		**Anger**	**Sad**	**Neutral**	**Happy**
	Anger	72.3	0	13.9	17
	Sad	6	84.4	1	12.8
	Neutral	0.9	0	99.7	0.7
	Happy	0	24.9	12.8	65.6

Comparably, the numerical matrix of conflict with the 1200 and 1500 epochs shows up in Table 5 and Table 6. Better estimates will be presented for all reactions in both tables. On the other tandem, happy and irritated individuals have come along. For the 1200 and 1500 epochs, the overall prediction accuracy is 78.4% and 80.5%, respectively. This result demonstrates that there is an improvement in prediction accuracy with an increase in training epochs [14].

5. Conclusion

The six groups that CNN utilizes to organize the entries include happiness, fear, sadness, disgust, frustration, and amazement. The network's performance is comparable to those of results reported in the scientific literature after it underwent training employing an ensemble of 200 speeches. FPGA circuits can be leveraged to implement the neural networks on hardware, but various internal monitoring exhibits can be utilized for validation of the circuits on-chip. Spectrophotograms gathered from the linguistic database have served as the models' input. Enhancing performance was found for the model when the overall number of training epochs elevated from 500 to 1200 and 1500. Due to the fact it does not include any advanced computational sciences concepts, it can be perceived as being easy to employ and understand. This work can be considered a practical tool for future work on the different appropriate sectors and will have significant consequences in the speech-emotion recognition field. It is, in real life, a cutting-edge method for utilizing neural networks and speech processing in practical situations to strengthen technology. Meanwhile, additional effort will be needed to further improve the current scheme to encourage the individual to acknowledge their emotions.

References

[1] Xhafa, F. (2017). Lecture Notes on Data Engineering and Communications Technologies.

[2] Pathar, R., Adivarekar, A., Mishra, A., & Deshmukh, A. (2019, April). Human emotion recognition using convolutional neural network in real time. In *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)* (pp. 1-7). IEEE.

[3] Alsobhani, A., ALabboodi, H. M., & Mahdi, H. (2021, August). Speech recognition using convolution deep neural networks. In *Journal of Physics: Conference Series* (Vol. 1973, No. 1, p. 012166). IOP Publishing.

[4] Deshmukh, R. S., Jagtap, V., & Paygude, S. (2017, June). Facial emotion recognition system through machine learning approach. In *2017 international conference on intelligent computing and control systems (iciccs)* (pp. 272-277). IEEE.

[5] Bai, X., Wang, X., Liu, X., Liu, Q., Song, J., Sebe, N., & Kim, B. (2021). Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. *Pattern Recognition*, *120*, 108102.

[6] Choudhary, R. R., Meena, G., & Mohbey, K. K. (2022, March). Speech emotion based sentiment recognition using deep neural networks. In *Journal of Physics: Conference Series* (Vol. 2236, No. 1, p. 012003). IOP Publishing.

[7] Vryzas, N., Vrysis, L., Matsiola, M., Kotsakis, R., Dimoulas, C., & Kalliris, G. (2020). Continuous speech emotion recognition with convolutional neural networks. *Journal of the Audio Engineering Society*, *68*(1/2), 14-24.

[8] Lakomkin, E., Zamani, M. A., Weber, C., Magg, S., & Wermter, S. (2018, October). 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.

[9] Ruiz-Garcia, A., Elshaw, M., Altahhan, A., & Palade, V. (2018). A hybrid deep learning neural approach for emotion recognition from facial expressions for socially assistive robots. *Neural Computing and Applications*, *29*, 359-373.

[10] Khan, A., Bahrami, M., & Anwar, Y. (2019, October). A Deep Learning Based Autonomous Mobile Robotic Assistive Care Giver. In *2019 IEEE International Conference on E-health Networking, Application & Services (HealthCom)* (pp. 1-8). IEEE.

[11] Tai, L., Li, S., & Liu, M. (2017). Autonomous exploration of mobile robots through deep neural networks. *International Journal of Advanced Robotic Systems*, *14*(4), 1729881417703571.

[12] Ristea, N. C., Duţu, L. C., & Radoi, A. (2019, October). Emotion recognition system from speech and visual information based on convolutional neural networks. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)* (pp. 1-6). IEEE.

[13] Alu, D. A. S. C., Zoltan, E., & Stoica, I. C. (2017). Voice based emotion recognition with convolutional neural networks for companion robots. *Science and Technology*, *20*(3), 222-240.

[14] Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Mansor, H., Kartiwi, M., & Ismail, N. (2020, September). Speech emotion recognition using convolution neural networks and deep stride convolutional neural networks. In *2020 6th International Conference on Wireless and Telematics (ICWT)* (pp. 1-6). IEEE.