Human Speech Emotion Recognition Using CNN

***Abstract*—Human speech emotion recognition is known to be the procedure of identifying emotion from the natural speech. It has become necessary to understand and detect the emotions of humans through various ways to provide a better user experience for the consumers. However, individuals have a wide range of diversity in their capacity of expressing emotions. Also, it may differ depending on how those emotions are being identified. Previously, different methodologies and techniques have been used to identify human emotions. An innovative convolutional neural network (CNN) based system for recognizing human speech and emotions is presented in this paper. Using a potent GPU, a model is constructed and fed with an unprocessed speech from a specified dataset for training, classification, and testing. the overall result was 94.38% which is quite better and surpasses many other models.**

***Index Terms*—Speech Emotion, CNN, Recognition.**

I. Introduction

With the advancements of rising smart devices and machine intelligence, the importance of emotion identification in human speech is growing dramatically as it has a significant impact on understanding the process of human decision-making. Also, it has been proven in several studies that learning this process can enhance the human-machine communication experience. It helps to enhance user experience in several cases by automatically acknowledging human emotional states. In recent years, human speech emotion recognition has attained a massive rise in demand for various applications. However, in human SER(speech emotion recognition) [1] it is frequently thought that audio data received from multiple devices can be merged into a single repository with a centralized approach, which is often unfeasible due to expanding data, bandwidth limitations, communication costs, and the absence of proper privacy concerns.

Many efforts have been made to detect human speech emotions keeping consumer privacy safe. However, to resolve this issue usage of federated learning attracted growing attention. Due to its distinctive flexibility to cooperatively train models of machine learning without sharing the local data and jeopardizing consumer privacy. Federated learning is especially well suited for this purpose. It is a setting for machine learning where multiple clients are seen to collaborate to train the models combined with a central server while keeping the data decentralized. In this environment, a central server is able to compile model updates from several clients during the training process. Clients can use federated learning to get the fundamental model that the server provides. This model may be seen as an ecosystem in which machine learning models and projects acquire data knowledge. Voice emotion identification is a task that may be used to identify the kind of emotion from human speech data. But the majority of recent research [2] that includes this environment has been on complete supervised learning procedures. However, finding supervised data [3] might be quite challenging for this scenario.

Recognizing human speech emotion recognition is a research area where many advancements have been made by applying several emerging techniques in recent years. With the advancements in technology as data collection has been easily attainable, Deep-learning has seen evolutionary growth as well. This way our speech emotion recognition research area has also made quite a usage of Deep-learning models like CNN, RNN, DSN, DNN, and so on. However, even though usage of RNN and other versions of RNN(such as LSTM) had benefits in training data and accuracy but such architectural models were seen to be increasing the computational cost. We intend on studying suitable data while reaching a higher accuracy and lesser computational cost. In this work, we are proposing a CNN approach for better performance of Human Speech Emotion Recognition.

II. Literature Review

In the research, paper [4] in order to achieve better outcomes, the wiener filter, a cascaded PRNN, and K-NN system, and a hybrid MFCC and GLCM system were each used in turn for each of these three phases. The Wiener filter was employed because it is simple to build and controls output error. When in a cascading system, PRNN and K-NN are used, a structure of the signal created by the emotional waves can be detected by the PRNN. Also, the nearest pattern of the signal is also to be more likely established by the K-NN method. This way, a hybrid system was developed that combines both MFCC and GLCM. This way GLCM's increasing usage of the spectrogram of sound which was generated by MFCC started to create a superior grayscale matrix for classification.

In the research, paper [5] federated learning was utilized to create the robot model. Following that, any client can train its own model and publish fresh prediction models to the server. When the operation is finished, the server will distribute the revised parameters once more after analyzing them. Each DTbot is a client that uses the same amount of patient data, trains the data locally, and only communicates with the server to update model parameters. The DTbot that was created did not need to send any images, videos, or audio files to the cloud for analysis because that kind of confidential material might include a lot of patients' in-depth private photos and discussions. It is crucial to safeguard privacy effectively while helping people in this day and age when it is highly valued. All of the hospital's robots have the capability to train their own personal data using the model provided by the host and communicate the results of any pertinent training

results directly to the host for further use. This ensures that the learning model will work and that the medical files will not need to be moved outside of the robot.

Guliani and his colleagues [6] trained voice recognition models using the decentralized approach of federated learning. FedAvg(federated averaging algorithm) and RNN-T architecture were utilized. The model is comprised of an LSTM audio encoder, an LSTM label encoder, a layer that concatenates the encoder outputs, and an output softmax. FedAvg was used to execute federated training of RNN-T models on a TensorFlow-based FL simulator running on TPU hardware. In this research study [1], Tsouvalas and his colleagues offer a federated learning-based SER model that protects user privacy. To remove the requirement of extensive labeled data availability on devices, researchers have used a data-efficient federated self-training technique to develop SER models using a small number of labeled data on devices. His team used ReLU as a non-linear activation function and an Adam optimizer with a learning rate of 0.001 to optimize categorical cross-entropy loss. Based on their examination, they revealed that the accuracy of their models regularly outperforms fully supervised federated settings with the same supply of labeled data. In this study, [3] Feng and Narayanan proposed a new Semi-FedSER framework that was presented to solve the constraints of limited labeled sample data in federated learning contexts for speech emotion recognition. Semi-FedSER utilizes labeled and unlabeled data samples in conjunction with pseudo-labeling at the local client. Semi-pseudo-labeling FedSER's strategy is based on multiview pseudo-labeling. His team also used the SCAFFOLD method to solve the problem of non-IID data distribution in the FL context. Results indicate that the proposed Semi-FedSER framework delivers accurate SER predictions despite the local label rate l = 20%.

In this research work [2], Tsouvalas and his colleagues investigate the practical challenge of semi-supervised federated learning for audio identification tasks. The customers have little to no motivation to classify their data, and for a number of crucial jobs, the subject expertise required to complete the annotation process effectively is lacking. On the other hand, vast quantities of unprocessed audio data are easily accessible on client devices. To address the dearth of class labels for learning on-device models, the authors describe a new self-training technique based on pseudo-labeling to utilize unlabeled on-device audio data and enhance the generality of models developed in federated environments. Regardless of its simplicity, they show that their technique, FedSTAR, is extremely practical for semi-supervised audio recognition training in a variety of federated setups and label availability. Comparing FedSTAR's performance with that of its fully-supervised, federated, conventional, and centralized equivalents, they conduct a comprehensive evaluation of FedSTAR on a variety of publically accessible datasets. The accuracy of the models regularly exceeds that of fully supervised, federated setups with the same label availability.

The acoustic modeling for children's ASR is the main topic of this paper [7]. In order to represent the short-term spectral envelope and capture information about the vocal tract system, standard short-term spectral feature extraction for speech recognition often uses a speech production model. The authors demonstrate that children's ASR systems can benefit from automatic feature learning by looking into one such strategy. They use the PF-STAR dataset for experimenting with children's speech and WSJCAM0 for adult speech. To conduct the experiment, WSJCAM0's standard training (train), development (dev), and test sets were employed. To decode WSJCAM0 utterances, the WSJ corpus' standard 20k trimmed trigram LMs were applied. The following is how the PF-STAR language model (LM) was created: Witten-Bell smoothing was used to build one LM using the training set, and Witten-Bell smoothing was used to build another LM using normalized text from the MGB-3 challenge. Also, they use CNN, GMM-HMM systems, DNN-HMM systems, and CNN-HMM systems. In comparison to their GMM/HMM and DNN/HMM counterparts, the CNN-based systems routinely outperform or are on a level with them. The SGMM systems also benefit from multi-pass decoding and data scarcity to produce accurate results. It is important to note that, to the best of our knowledge, the performance reported at 11.99% WER is the best on the PF-STAR corpus.

In the research paper [8], they developed a multi-task framework for their suggested method, whereby they jointly train an accent classifier and explicitly supervise a multi-accent acoustic model with accent information. Additionally, they train a different network that learns accent embeddings, which can be included as auxiliary inputs within our multi-task architecture. They decided to create our acoustic model using time-delay neural networks (TDNNs). Also, the Common Voice corpus from Mozilla was used for this experiment. Common Voice is a corpus of reading speech in English that is crowd-sourced from a large number of speakers residing in different parts of the world. In this study, the authors investigate the application of a multi-task architecture for accented speech recognition, where a multi-accent acoustic model is concurrently learned with an accent classifier. In comparison to a multi-accent baseline system, this network performs significantly better, reducing WER by up to 15% on a test set with visible accents and by 10% on unnoticed accents. Performance is further enhanced by accent embedding acquired from a separate network.

In the research paper [9] the fundamental method put out is based on two separate vectors, suggesting that a mood will be produced based on where two vectors are situated on the emotion planes. Three main sorts of techniques can often be used by modern speech processing systems. The first of them takes into account certain spectral characteristics. In this case, the speaker's general features can be revealed without regard to any particular phoneme characteristics. The second method uses feature vectors for a quick training phase. Regrettably, the amount of training vectors required for real-time emotion recognition is so high that it exceeds the memory and processing power of current computers. As a result, it's required to use some unique solutions, such as vector quantization (VQ)

strategies or HMM-based approaches. A VQ codebook is made up of a limited amount of straightforward yet very specific feature vectors. By grouping these vectors according to this codebook, it is feasible to reflect certain speaker attributes. The third technique uses speech recognition techniques. Since various languages pronounce the same phoneme differently, phoneme templates developed through training processes can be used to identify emotions.

The authors of this article [10] suggest a methodology for SER by using a DSCNN(Deep Stride Convolutional Neural Network) based framework that is mostly used in plain nets in terms of Computer Vision tasks. It has been implemented through scikit-learn packages. They employed transfer learning strategies to train the AlexNet, Vgg-16, and Resnet-50 CNN's models using the IEMOCAP dataset. Via this model, the accuracy of 81.75% has been achieved

III. Datasets

One is RAVDESS Emotional speech audio [11] which consists of 1440 files with randomized data of speech collected from songs, audio, and videos. Among 1440 file data there are 60 trials per actor. 720 files are of Female voices and the other 720 are of Male voices. This data consists of 7 identifiers that are, Modality, Vocal channel (01 = speech, 02 = song), Emotion, and Emotional intensity. Statement, Repetition, Actor Number. This dataset includes 3 sorts of modalities which are fully Audio—Video, only Audio, and only Video. The speech is either retrieved from an original speech of a song. It has 8 classified emotions which are labeled by numbers. In this dataset, the Neutral state is denoted as 01, Calm as 02, 03 indicates happy state emotions, 04 points to sad emotions, number 05 resembles angry emotion, 06 is fearful, disgust is labeled 07 and lastly, 08 denotes surprising emotions. However, This dataset is unable to classify neutral emotional states. It has only 2 types of emotional intensity that are normal and strong. The actors spoke a selection of only 2 sentences.

Dataset CREMA-D [12] focused more on accents and pronunciations by choosing actors of different ethnicity. It consists of 7442 clips from 91 different actors. There was a selection of 12 sentences from which the chosen actors spoke their parts. This dataset consists of 6 different emotions that are Anger, Disgust, Fear, Happy, Neutral, and Sad. Also, all these classes have can be again classified in 4 different levels through which emotions can be expressed. These are, Low Level, Medium Level, High Level, and Unspecified.

Toronto emotional speech set (TESS) dataset [13] has a collection of 2800 data where they set 200 target words recorded by only 2 female actors. This dataset is portraying 7 emotions which are, anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral.

SAVEE-dataset(Surrey Audio-Visual Expressed Emotion) [14] was initially recorded from 4 English male speakers who wear also native English speakers. This dataset is portraying 6 different emotions that are, anger, disgust, fear, happiness, sadness, and surprise. However, this dataset also added a neutral category. In text material consisted in this dataset, per emotion, there are15 TIMIT sentences. Also from there, 3 are commons, 2 of those are emotion-specific, and 10 are known to be generic sentences that were different for each emotion and phonetically balanced.

IV. Methodology

A. Data preprocessing :

As we have four different datasets: Crema-D, Ravdess, Savee, Tess. Therefore, we must establish a data frame that stores all emotions of the data together with their pathways. This data frame will be used to extract features for our model's training. As for the datasets we need to extract files for each audio according to their emotions and encode integers to actual emotions. Then we have to save all the datasets to a file path. The primary component of a speech emotion recognition system is feature extraction. It is performed mostly by transforming the voice waveform into a parametric representation at a comparatively lower data rate. In feature extraction, we have used zero crossing rate, and MFCC (Mel Frequency Cepstral Coefficients).

B. Audio Augmentation :

We can construct syntactic data for audio by injecting noise, altering time, and modifying pitch and tempo. We used noise figure 1, stretched, shifted, and pitch for audio augmentation.

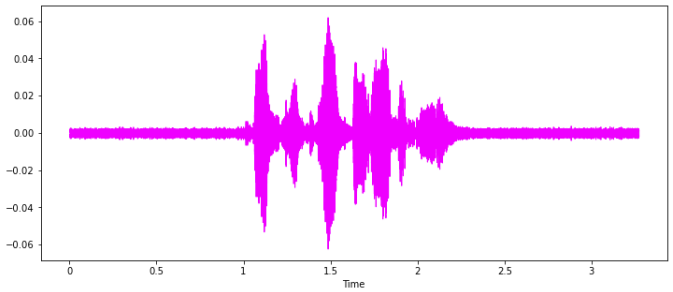

Fig. 1. Noise added in audio

C. Architecture of Deep Convolutional Network :

This Model has a total of 6 one-dimensional convolutional layers as we can see from figure 2, each followed by a batch normalization layer and a max pooling layer with a pool size of 2, except the last Conv1D which has a pool size of 3. The number of filters in the convolutional layer is 512, 521, 256, 256, and 128, followed by the kernel size of 5, 5, 5, 3, 3. The activation function of the 1st Dense layer is 'Relu' and the second one we used is the softmax function.

D. Training CNN model

Because our audio characteristics have a time dimension, we chose a one-dimensional convolutional neural network. The time-based structure of an audio wave is utilized by the 1D CNN kernels linear progression. The key fundamental blocks of CNN are the convolutional layer, the pooling layer, and

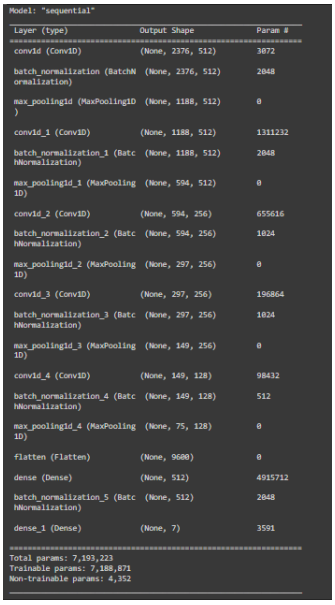

Fig. 2. CNN model

the fully connected layer. In order to train the model, we used 'adam' optimizer with an initial learning rate of 0.00001. The loss function is used because it measures how good the prediction model does in terms of being able to predict the expected outcome. For the loss function, we used 'Categorical cross-entropy. The model trained over 50 epochs.

V. Results

The previous model of CNN as it is shown in figure 3 used 4 one-dimensional convolutional layers. The number of filters for convolutional layers are 256, 256, 128, and 64. Followed by the kernel size 5, 5, 5, 5. We also used 2 dense layers for this model. We can clearly see from figure 4 that the accuracy of our model on test data was around 60.74% .

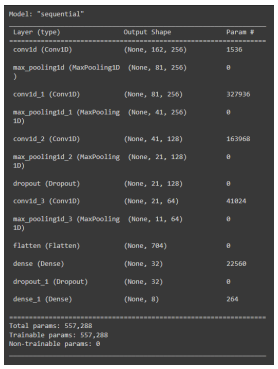

Fig. 3. 4 layers CNN model

By adding some extra features for data augmentation techniques and using other feature extraction methods and also creating extra two layers for convolutional layers along with increased filter size we have achieved the accuracy 94% for our test data as we can see from figure 5.

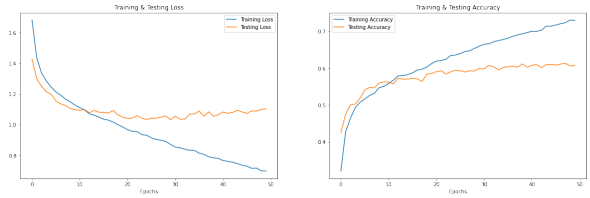

Fig. 4. Training and Testing accuracy 4 layers CNN model.

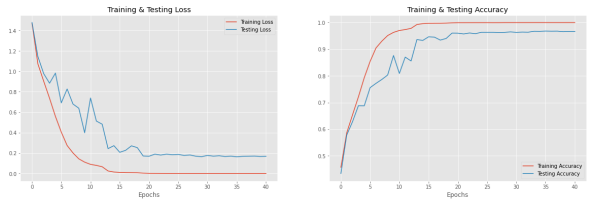

Fig. 5. Final model Training and Testing accuracy.

Moreover, we can see from the confusion matrix in figure 6, that our model is predicting every features quite well.

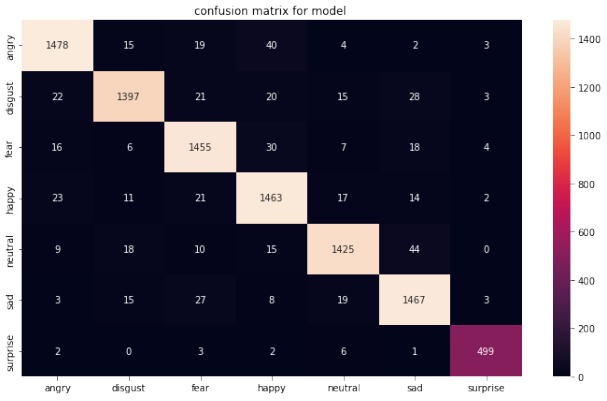

Fig. 6. Confusion matrix

Our model performs best when predicting shock and anger, which makes sense given the wide variety of differences between audio files containing expressions of such moods and those containing neutral expressions. Ultimately, we were able to obtain 94% accuracy on our test data as we can see from figure 5, which is respectable but still has room for improvement through the use of additional augmentation approaches and alternate feature extraction strategies.

From the above table I we can clearly see that our model is predicting angry 95% of the time, disgust 96%, fear 94%, happy 93%, neutral 95%, sad 93% and the best is predicting surprise which is 97%.

VI. Conclusion

Human speech emotion recognition is a field where many researchers are working to develop a system that is capable of comprehending the state of a human voice to assess or detect the speaker's emotional state. The human speech

Emotions	Precision	Recall	f1-score	support
Angry	0.95	0.95	0.95	1561
Disgust	0.96	0.93	0.94	1506
Fear	0.94	0.95	0.94	1536
Happy	0.93	0.94	0.94	1551
Neutral	0.95	0.94	0.95	1521
Sad	0.93	0.95	0.94	1542
surprise	0.97	0.97	0.97	513
Accuracy			0.94	9730
Macro avg	0.95	0.95	0.95	9730
Weighted avg	0.94	0.94	0.94	9730

TABLE I
EVALUATION METRICS OF THE MODEL

emotion recognition literature confronts many difficulties to increase recognition precision while reducing the computing complexity of the model. To resolve those issues we have applied a CNN(Convolutional Neural Network) architecture while improving the dataset by merging Four different datasets basing on it's classes. After successful application of our model and by analyzing achieved results it is a matter of fact that the model has given better accuracy compared to the results achieved in many other previous works that we reviewed in earlier sections. However, as we previously discussed the recent growing concerns about keeping user privacy safe in such tasks, unfortunately, our model doesn't take this issue into the account. In the future, we would like to resolve this issue through the usage of federated learning in our work. As a consequence of utilizing federated learning, we intend on training our Deep Learning architecture model without sharing our consumer's data locally later on.

REFERENCES

[1] V. Tsouvalas, T. Ozcelebi, and N. Meratnia, "Privacy-preserving speech emotion recognition through semi-supervised federated learning," in *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, 2022, pp. 359–364.

[2] V. Tsouvalas, A. Saeed, and T. Ozcelebi, "Federated self-training for semi-supervised audio recognition," *ACM Transactions on Embedded Computing Systems (TECS)*, 2021.

[3] T. Feng and S. Narayanan, "Semi-fedser: Semi-supervised learning for speech emotion recognition on federated learning using multiview pseudo-labeling," *arXiv preprint arXiv:2203.08810*, 2022.

[4] J. Umamaheswari and A. Akila, "An enhanced human speech emotion recognition using hybrid of prnn and knn," pp. 177–183, 2019.

[5] Y. Liu and R. Yang, "Federated learning application on depression treatment robots (dtbot)," in *2021 IEEE 13th International Conference on Computer Research and Development (ICCRD)*. IEEE, 2021, pp. 121–124.

[6] D. Guliani, F. Beaufays, and G. Motta, "Training speech recognition models with federated learning: A quality/cost framework," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3080–3084.

[7] S. P. Dubagunta, S. H. Kabil, and M. M. Doss, "Improving children speech recognition through feature learning from raw speech signal," pp. 5736–5740, 2019.

[8] A. Jain, M. Upreti, and P. Jyothi, "Improved accented speech recognition using accent embeddings and multi-task learning." pp. 2454–2458, 2018.

[9] Z. Ciota, "Emotion recognition on the basis of human speech," pp. 1–4, 2005.

[10] S. Kwon, "A cnn-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, p. 183, 2019.

[11] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[12] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[13] K. Dupuis and M. K. Pichora-Fuller, "Toronto emotional speech set (tess)-younger talker_happy," 2010.

[14] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," 2014.