Emotion Detection and Voice-Emotion Conversions Using Deep Learning

Yeshitha B
Electronics and Communications
R V College of Engineering
Bengaluru, India
yeshithab.ec19@rvce.edu.in

Vinitha V
Electronics and Communications
R V College of Engineering
Bengaluru, India
vinithav.ec19@rvce.edu.in

Anubha Mittal
Electronics and Communications
R V College of Engineering
Bengaluru, India
anubhamittal.ec19@rvce.edu.in

P. Harshitha Reddy
Electronics and Communications
R V College of Engineering
Bengaluru, India
ponkalaharshitareddy.ec19@rvce.edu.in

Rajani Katiyar
Electronics and Communications
R V College of Engineering
Bengaluru, India
rajanikatiyar@rvce.edu.in

***Abstract*— Emotion, especially through speech, is a powerful tool humans possess that conveys much more information than any text can describe. Using artificial intelligence to tap into this can have a big positive impact on a variety of industries, including audio mining, customer service applications, security and forensics, and more. A growing field of research, spoken emotion recognition, has relied heavily on models that employ audio data to create effective classifiers. This paper presents convolutional neural network as a deep learning classification algorithm to classify 7 emotions with an accuracy of 69.45% on the combined datasets of Savee, Ravdess and Tess. It proposes a new system to help replicate the emotions on a neutral audio (voice conversion). The production of the emotional audio is implemented using MelGan, a special type of Generative Adversarial Network (GAN).**

Keywords— ser, mfcc, asr, SVM, CNN, MLP, GAN, Mel, MelGAN, Generator, Discriminator

I. INTRODUCTION

One of the most fundamental and important capacities possessed by humans is spoken communication. The single most essential way that people can easily share knowledge without the use of any "carry-along" tools is through speech. Although our eyes acquire more external cues than our ears do on a passive basis, visual communication is almost entirely worthless when compared to verbal communication. The speech wave itself carries linguistic information, the speaker's voice traits, and their emotional state. The linguistic and acoustical components of speech have, and continue to have, a close relationship with our cultural and social evolution. The speech wave transmits various types of information, primarily linguistic information that shows the meaning the speaker wants to convey, individual information that identifies the speaker, and emotional information that expresses the speaker's feelings.

Emotion recognition is crucial in many intelligent interfaces [1]. This is still a challenging problem despite recent developments in deep learning (DL). The fundamental issue is that DL models are prone to over-fitting since the majority of publicly accessible annotated datasets in this field are modest in scope. The fact that we communicate our feelings in a variety of ways is another crucial aspect of emotion recognition [2]. Numerous techniques, such as the analysis of EEG, body postures, and facial expressions, can be used to detect emotional information [3]. Speech is likely the most approachable of them. Speech signals include a variety of different emotional indicators in addition to accessibility [4]. Despite the fact that speech signals contain a significant amount of information, it can be unsatisfying to remove the linguistic component that coexists with it. This is especially true considering that the text component can be easily transcribed in practical applications thanks to the notable successes in the speech-to-text domain and the availability of several commercial-scale APIs [5].

II. LITERATURE SURVEY

In this section summarizes some of the research in the field of Speech Emotion Recognition (SER). The task of SER is not new and has been studied in the literature for quite some time. Previously for speech emotion recognition the classical machine learning models like Hidden Markov model (HMM), Support vector machine (SVM), and decision trees were used. In order to improve the efficiency several neural network-based architectures have been proposed. Deep neural networks, which were recently introduced into the domain, have also significantly improved state-of-the-art performance. In order to demonstrate the effectiveness in SER an early study has used deep neural networks (DNNs) and extracted high-level features from the raw audio data. The advancement has increased in more complex neural-based architectures. Convolutional neural network (CNN) models were trained on data extracted from the raw audio signals using spectrograms. The audio features include Mel-frequency cepstral coefficients (MFCCs) and low-level descriptors (LLDs). Another area of study has been the use of different machine learning methods in combination with neural network-based algorithms. One of the researchers had used the multi-object learning approach and auxiliary tasks such as gender and naturalness to help the neural network based model in order to learn various features from the given dataset [9]. Another researcher has looked into transfer learning methods that used the external data from the domains which are related. [10].

It was also observed that increasing the dataset results in better training of the model and therefore a higher overall accuracy [15]. While analyzing acoustic features of speech with the help of discrete classification of emotions, three main features are extracted that are MFCC, Mel Spectrogram and Chroma. Different algorithms, such as SVM, XHB, CNN-1D(Shallow) and CNN-1D &2D (Deep), were evaluated with different configurations and input features in order to design a novel hierarchical technique for emotion

classification using 5 training datasets. The ensemble of CNN-2D(deep) and CNN-1D (shallow) and CNN -1D (deep) showed the best results.

The effects of different activation functions were studied for SER systems on a LSTM model. It was observed that the disadvantages of RNN's vanishing gradient problem could overcome by using long short-term memory. Finally, SER system's calculated accuracies performed admirably on crema-D dataset. The proposed model analyses the end results using three different activation functions which are Sigmoid function, Tanh function and Rectified Linear unit. Gradient descent is used as the basic optimizer in linear equation to convert the loss in previous layer as a weight in the next. With this proposed LSTM recurrent network, the efficiency got using activation function relu (0.588389246) is higher than tanh (0.552054216) and sigmoid (0.463087260) functions. Because of this RELU function is considered to be computationally efficient.

In another study were MLP classifier was applied on the Ravdess dataset due to its ability to handle highly volatile data having non-constant variance [18]. Taking mfcc and chroma as input features for MLP classifier the efficiency obtained is 82.86%. Speech emotion recognition method which uses the combination of both acoustic and language adaptation model. To extract linguistic features speech is converted into text by using high -accuracy ASR. eVectors features are used for extracting emotional features from text for emotion recognition. For emotion recognition by acoustic features, low-level-descriptor LLD are extracted and given as inputs for DNN layer. When both the acoustic and linguistic features were used, it increased the performance and accuracy is 77.25%.

Automation is the future. Every system is being upgraded to provide for interactive interfaces. In regard to this, the current technology only can communicate in a neutral plain voice. With emotion in them, these systems will have wider application range as well better effectiveness in serving man. GANs are strong generative models that simultaneously train two competing networks, a generator and a discriminator, in an effort to approximate the data distribution [19]. Numerous studies have concentrated on raising generated sample quality and stabilizing GAN training [20, 21]. Recently, data augmentation has made use of the GAN's capacity to provide accurate in-distribution samples. In [22], the authors train a GAN to provide in-class samples. The Cycle GAN architecture [21] is modified in [23] for the classification of emotions from facial expressions. In [24], artificial feature vectors are employed to enhance the classifier's performance on an emotion task in the voice domain. To solve the data imbalance, a conditional GAN design is suggested in [25].

In this paper, detecting emotions from speech and voice conversion by imposing an emotion on a neutral audio is implemented. This is done by first segregating emotions from the input audio signals. The input is acquired from the datasets Savee, Ravdess and Tess. This amounts to 5000 data samples. CNN deep learning model using spectrograms of the audio signal is used to classify the emotions. Then, the emotion has been replicated on a neutral signal to produce the voice converted signal.

III. METHODOLOGY

In this work, the effect of different models Both ML and DL on SER is studied to classify seven emotions; happy, sad, fear, disgust, surprise, angry and neutral. The ML models include SVM (support vector machine) and decision tree while DL models are MLP, LSTM and CNN.

A. DATASET

Three different Databases namely The Ryerson Audio Visual Database of Emotional Speech and Song (RAVDESS), Toronto emotional speech set (TESS), and Surrey Audio-Visual Expressed Emotion (SAVEE). All the audio files are taken in .WAV file format. In total there are 4528 audio files all put together.

1) RAVDESS: Contains 24 professional actors (12 female, 12 male), speech emotions include calm, happy, sad, angry, fearful, surprise, and disgust expressions. There are 1440 audio files.

2) TESS: Contains seven emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. There are 2800 data points (audio files) in total.

3) SAVEE : Conatains english male speakers (identified as DC, JE, JK, KL). Emotions are anger, disgust, fear,happiness, sadnesss and surprise. There area total of 480 audio files.

The datasets are divided into testing and the training datasets. The model is trained using the training dataset and is tested using the testing dataset.

The dataset is split into training and testing. The training set is used to train the model and the test set is a subset of the training set but is not used in training rather it is used to validate the model performance after training is completed. A 90-10 split is used in this project, that is 90% training and 10% for testing. Out of 4528 audio files 4075 files are used for testing and 453 for testing.

B. SPLITTING OF DATA

The dataset is split into training and testing. The training set is used to train the model and the test set is a subset of the training set but is not used in training rather it is used to validate the model performance after training is completed. A 90-10 split is used in this project, that is 90% training and 10% for testing. Out of 4528 audio files 4075 files are used for testing and 453 for testing.

C. FEATURE EXTRACTION

Frequencies are the important features that needs to be extracted from the audio to train the datasets for speech emotion detection. Techniques like FFT (Fast Fourier Transform), STFT (Short Time FFT), MFCC (MELFrequency Cepstral Coefficients) are used.

1) FFT (Fast Fourier Transform): It converts the time domain audio signal to frequency domain which represents the power occupied in each frequency of the audio signal hence this results in losing the time aspect of the audio signal. Therefore, FFT results in static nature of frequency extraction whereas the audio data changes with time.

2) STFT (Short Time FFT): To overcome the limitation of FFT that is, FFT results are static to determine the emotion of the audio file, several Fourier transforms are taken at different time intervals. Hence this holds information on time as well as on frequencies.

3) MFCC (Mel-Frequency Cepstral Coefficients): MFCC feature extraction results in better emotion detection as MFCC approximate the functioning of a human ear to distinguish between different melodies when it has same frequency and rhythm. Hence this feature gives output as perceived by human beings. It imitates or mimics cochlea of human ear. MFCC is inverse FFT of log of spectrum. Hence this feature is used for feature extraction

IV. MODELS AND RESULTS

1) SUPPORT VECTOR MACHINE (SVM):Datasets are classified as testing and training datasets. SVM finds a plane of separation between different datapoints as there are 7 different labels for the emotion detection SVM is not very efficient in finding a plane of separation between all the emotions. RBF (Radial Basis function) kernel is used for training phase. The main advantage of RBF is that it restricts training data to lie in specific boundaries. This kernel will map the samples into a higher dimensional space. Machine Learning models are not capable of training huge number of datasets hence the training accuracy is 15.78% and the testing accuracy is 13.79% (Table 1).

2) DECISON TREE (DT): Decision tree classifies the input data by using a hierarchy of tree structure which is split based on entropy or Gini-index method which signifies the most important attributes which are necessary to classify the input in a hierarchical way. It reduces the confusion between emotions. Since the features or attributes used is only MFCC which is an array of values it inconclusive to choose the most important attributes out of all the numbers in the array. In this paper Gini-Index is used to find the most influencing attribute. When the model is run and tested the accuracy for training is 46.44% and for testing is 44.26%. (Table 1)

3) MULTI-LAYERED PERCEPTRON (MLP): The input layer in MLP accepts the input features, which are then transmitted through numerous hidden layers to produce the classification result at the output layer. It is based on a feed forward network. The inner nodes in the hidden layers use nonlinear activation functions to classify the emotions. An adaptive learning rate is used with 10000 hidden layers, and max_iter being 500 the testing accuracy is 19.19% and the testing accuracy is 16.28%. (Table 1).

4) Long Short-Term Memory (LSTM) : LSTM model helps in solving time series forecasting problems. It learns from the past observations and predicts the next sequences. As audio signals are not very predictable in nature the use of LSTM model is less significant for this application. It has a feedback connection which helps in processing the audio sequences. This model will get the input from cells and manipulates it using gates. There are three gates:

a) Forget Gate(f): It helps in forgetting the previous data.

b) Input Gate(i): It helps in determining what has to be input.

c) Output Gate(o): It helps in determining the output which has to be sent. In this paper a 3-layer LSTM Network is used with activation function as relu and at the output later activation function used is SoftMax. After the evaluation of the model the training accuracy achieved was 16.34% and testing accuracy was 12.9%.

5) CONVOLUTION NEURAL NETWORK: A CNN (Convolutional Neural Network) is a type of artificial feed-ahead network where the joining sequence between its nodes is prompted by the presentation of an animal visual-cortex. Computationally, convolution procedures can be used to mathematically reproduce the cortical neurons that respond to signals in a single node's receptive area. CNN is an extension of multi-layer perceptron designed to utilize very little pre-processing. The model used consists of 4 neutral network layers, having 2 Conv1D convolution layers, 2 dense layers with a rectified linear unit (Relu) and using nadam optimizer. The dropout of 0.2 used in the network to reduce the system overfitting. The model is trained with batch_size=16 and 13 epochs. When the model built is tested using the testing dataset the training accuracy is 69.45% and the testing accuracy is 67%.

In paper [26] the accuracy obtained by using CNN on IEMOCAP dataset is 65.35%, in this paper it has achieved an accuracy of 67%.

```
[[125   0   0   0   0   0   0]
 [140   0   0   0   0   0   0]
 [135   0   0   0   0   0   0]
 [121   0   0   0   0   0   0]
 [140   0   0   0   0   0   0]
 [116   0   0   0   0   0   0]
 [129   0   0   0   0   0   0]]
```

Figure 3, *Confusion Matrix for SVM*

The rows in the confusion matrix represent the output class and the columns represent the target class. From Fig 3 we can say that angry has 125 samples correctly predicted as angry and 140 samples wrongly predicted as disgust, 135 samples wrongly predicted as fear, 121 samples wrongly predicted as happy, 140 samples wrongly predicted as neutral, 116 samples wrongly predicted as sad, 129 samples wrongly predicted as surprise. Hence it can be concluded that the target class is always been angry and the model is predicting for all output classes hence signifying that emotion angry is getting confused with all other emotions and only one emotion is considered as a target class. This gives us a poor result.

```
[[85  5 14  8 10  1  2]
 [28 68  5  5  8 16 10]
 [31  4 65 14  2  7 12]
 [29  9  6 36 18  7 16]
 [31  5  7 20 58 11  8]
 [29 13  0  6  5 59  4]
 [31  9  4 23  5  8 49]]
```

Figure 4, *Confusion Matrix for DT*

In Fig 4 we can conclude that Decision tree performs better than MLP,SVM and LSTM as the model is predicting

for all the target classes and also the diagonal elements (which gives the number of examples correctly predicting that it belongs to the same class) are also in good numbers like 85 samples are correctly predicted as angry, 68 samples correctly predicted as disgust, 65 samples correctly predicted as fear , 36 samples correctly predicted as happy, 58 samples correctly predicted as neutral, 59 samples correctly predicted as sad and 49 samples correctly predicted as surprise. Hence it can be concluded that this model has predicted surprise emotion very well than SVM, LSTM and MLP.

```
[[ 0 14 59 32  0 20  0]
 [ 0 21 31 27  0 61  0]
 [ 0  2 86  5  0 42  0]
 [ 0  5 48 36  0 32  0]
 [ 0  0 45 69  0 26  0]
 [ 0  7 17 57  0 35  0]
 [ 0  2 65 32  0 30  0]]
```

Figure 5, *Confusion Matrix for MLP*

In Fig 5 we see that 1st column,5th column and 7th column are zeros which means that emotion angry, neutral, and surprise have not been predicted at all by the model. We also observe that fear has 86 samples correctly predicted as fear, but 48 samples wrongly predicted as happy and 59 samples wrongly predicted as angry hence it can be concluded that emotion fear is getting confused with happy and angry. This is one inference that could be drawn by using MLP model. Similarly other conclusions can be made by taking other emotions into consideration.

```
[[  0   0   0   0   0 125   0]
 [  0   0   0   0   0 140   0]
 [  0   0   0   0   0 135   0]
 [  0   0   0   0   0 121   0]
 [  0   0   0   0   0 140   0]
 [  0   0   0   0   0 116   0]
 [  0   0   0   0   0 129   0]]
```

Figure 6, *Confusion Matrix for LSTM*

Fig 6 also has the same results but the target class here is sad. Hence here sad is getting confused with all other output classes.

```
[[ 83   5  12  14   6   2   3]
 [  6  91   8   6   9   9  11]
 [  6   3  86   5   6  10  19]
 [  5   7  10  64  10   2  23]
 [  3   5   3   9 105   7   8]
 [  1  12   6   2  16  73   6]
 [ 10  10  13   9   6   3  78]]
```

Figure 7, *Confusion Matrix for CNN*

In Fig 7 we can see that CNN model also performs well in predicting for all the output classes. As compared to Confusion matrix of DT In Fig 5.4 we can conclude that this model outperforms all other models as angry has only 6 samples wrongly predicted as disgust and fear, 5 samples wrongly predicted as happy, 3 samples wrongly predicted as neutral, 1 sample wrongly predicted as sad and 10 samples wrongly predicted as surprise these numbers are very much lesser that the values seen in 1st column of Fig 5.4. It can also be inferred that emotion surprise is predicted better than the results seen in DT. The diagonal element values are higher than the diagonal element values in Confusion matrix of DT hence inferring that the number of samples/examples correctly predicting that it is belonging to the same class is higher that DT results.

6) F1 Score: F1 score talks about both precision as well as recall, TABLE-3.

a) It is seen that SVM and LSTM models perform very poor in detection of emotions this is because SVM works with the principles of a hyperplane dividing the datapoints in the space. As the number of classes are more hence it is very inefficient in determining the emotion whereas LSTM works better with text sentences which takes into consideration about the previous input and then determine the output but audio signals as such are very unpredictable hence failing to detect the emotion.

b) Decision Tree outperforms SVM, LSTM and MLP. It can be seen that happy has least F1 score. The emotion happy can be misinterpreted as angry, disgust, or fear hence leading to less F1 score for happy. Neutral can be confused with sad. Though it performs better than SVM, LSTM, and MLP its overall performance is not appreciable. MFCC contains a set of vectors hence it is difficult in determining the most important vector in deciding the emotion of the audio.

c) MLP fails to predict for angry, neutral and surprise. The F1 scores of the rest of the emotions are also very low. It performs better than SVM as it is capable to learn non-linear models.

d) CNN outperforms all the other models. The F1 scores of all the emotions are balanced hence proving that the emotions are not getting misinterpreted or confused as it happened in DT.

7) GANs: For unsupervised learning, neural networks called Generative Adversarial Networks (GANs) are employed. It is composed of a system of two competing neural network models, the Generator and Discriminator, which are capable of identifying, capturing, and copying the variations present in a dataset. This method develops the ability to produce fresh data. A random noise vector serves as the generator's input. Real datasets and fake data produced by the generator are used as input to the discriminator.

a) Generator: The Generator creates fictitious data samples in an effort to deceive the Discriminator. By tricking, it seeks to increase the chance that the Discriminator will fail.

b) Discriminator: The discriminator aims to differentiate between authentic and bogus samples. that calculates the likelihood that the sample it received came from training data rather than the generator. When the discriminator successfully identifies real and fake samples, it is rewarded, whereas the generator is penalized with large updates to model parameters and vice versa when generator is penalized. The voice conversions based on this model gives satisfactory results in imposing the emotion on a neutral audio signal. The output audio is represented as spectrograms. The model was run for 50 epochs and a significant improvement is seen between the first epoch (figure 8) and last epoch (figure 9).

8) MelGAN: MelGAN is a sub-type of GANs that works best for audio systems. It is faster and uses lesser hardware while being highly efficient without hampering the audio quality. It consists of the generator block and discriminator block.

a) Generator: The generator block takes in a Mel spectrogram as its input as opposed to random noise in GAN. It has two up sampling layers that has dilated convolution between them. It is not necessary to produce samples one at a time in MelGAN. The association between each sample pair in MelGAN is implicit rather than directly causally dependent. Even temporally distant samples can share a lot of input mel-spectrogram frames and hidden layer nodes by employing a lot of dilated convolution blocks, but the samples are not directly dependent on one another. As a result, MelGAN allows for total parallelization of audio sample production.

b) Discriminator: In this paper 3 discriminators have been used. Each discriminator operates on a window basis that is one discriminator works within one widow frame of the whole audio sequences. Each frequency component in the discriminator is independent of the other discriminator. By applying this window technique and using three discriminators it can be ensured that the random noise generated by the generator is captured by either one of the discriminators hence the error introduced in easily captured by the discriminator and equalizing the errors. Thus, reducing the burden on one discriminator which works on the whole audio sequence as used in GANs. By using MelGANs this paper has successfully translated neutral emotion to happy and sad. It could also incorporate both emotion as well as voice conversions. Conversion from Male-Neutral to Female-Sad and Female-Angry. The spectrogram results are in Figure 10, 11 and 12.

In paper [27] Voice conversions on English and Urdu Corpus which contains single word utterance from each speaker using Cyclic GANs is carried out. This project has achieved voice along with emotion conversions for a window frame of 1-2 seconds by using MelGANs.

TABLE 1. *Testing and Training accuracy of respective models*

MODEL	TRAINING ACCURACY	TESTING ACCURACY
Support Vector Machine (SVM)	15.78%	13.79%
Decision Tree (DT)	46.44%	44.26%
Multilayer Perceptron (MLP)	19.19%	16.28%
Long Short-Term Memory (LSTM)	16.34%	12.9%
Convolutional Neural Networks (CNN)	69.45%	67%

TABLE 2. *Results obtained from models for the respective epochs*

NO. OF EPOCHS	MLP	LSTM	CNN
5	9.08%	8.32%	32.41%
10	11.23%	10.65%	40.04%
15	14.56%	12.58%	48.78%
20	12.34%	13.33%	59.97%
25	16.28%	12.9%	67%

TABLE 3. *F1 score of all 7 emotions for different models.*

MODEL	ANGRY	DISGUST	FEAR	HAPPY	NEUTRAL	SAD	SURPRISE
SVM	0.242483	0	0	0	0	0	0
DT	0.43701	0.5375	0.55084	0.3090	0.47154	0.5244	0.426086
MLP	0	0.2198	0.35390	0.18997	0	0.19327	0
LSTM	0	0	0	0	0	0.227	0
CNN	0.69456	0.6667	0.63	0.5565	0.7469	0.65765	0.56317

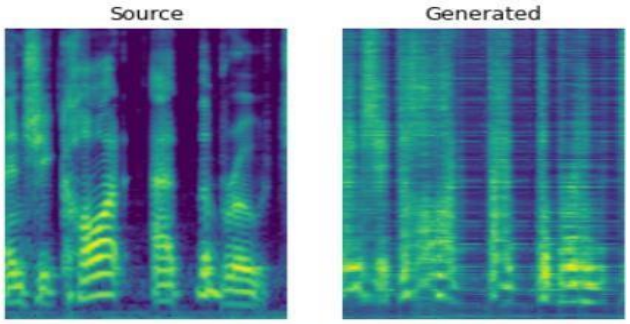

Figure 8, *Spectrogram result obtained in first epoch of MelGan implementation*

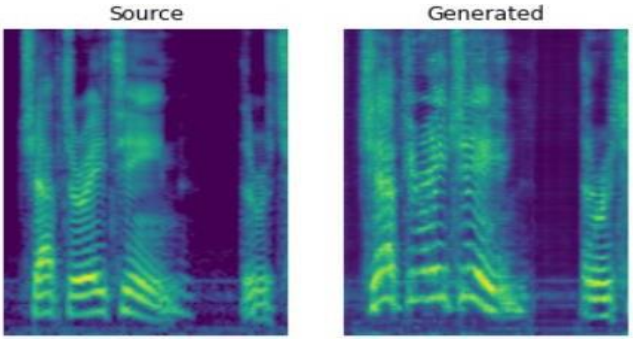

Figure 9, *Spectrogram result obtained in the 50th epoch of MelGan implementation*

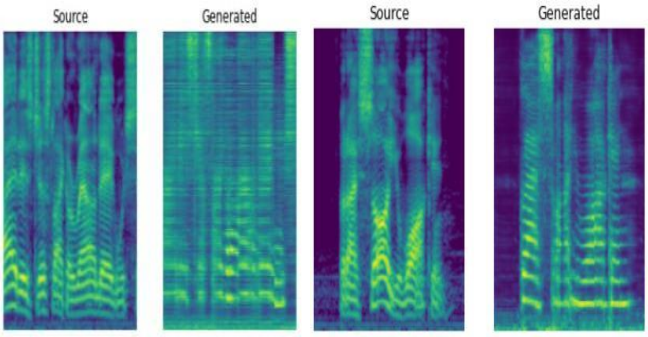

Figure 10, *MelGAN output spectrograms for male neutral to sad-female conversion of epoch 1(left) epoch 50 (right*).

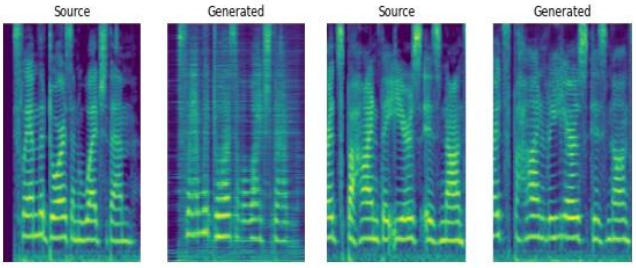

Figure 11, *MelGAN output spectrograms for female neutral to female- sad conversion of epoch 1(left) epoch 50 (right).*

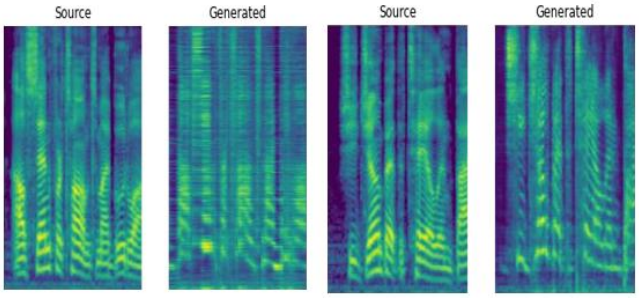

Figure 12, *MelGAN output spectrograms for male neutral to female-angry conversion of epoch 1(left) epoch 50 (right)*

V. CONCLUSION

This paper has provided a detailed review of various machine learning and deep learning techniques for SER. It is concluded that CNN gives the best results among all the models. MFCC of the speech signal serve as the input to deep CNNs. CNN consists of three convolutional and three fully connected layers, and predictions for the seven emotion classes are produced. The F1 scores of all the emotions are balanced hence proving that the emotions arenot getting misinterpreted or confused as it happened in DT. The proposed framework has to be improved in order to recognize all emotions robustly and effectively. More data combined with moderately complicated models can boost SER performance even more for future scope.

Voice Conversions helps in converting the voice from male to female or vice versa at the end of 50th epoch it is very distinguishable that the male audio is converted to female audio by preserving its linguistic contents also. This concept is applicable in devices such as Alexa and Siri. The MelGan model efficiently converts the neutral signal to emotion. The best results were obtained for successful conversion from neutral emotion to happy and sad. It could also incorporate both emotion as well as voice conversions. Conversion from Male-Neutral to Female-Sad and Female-Angry.

However, emotion like sarcasm is a field still unexplored hence the future scope of this paper lies in working towards sarcastic statements which are important in psychological aspects.

REFERENCES

[1] R. Ezhilarasi, R.I. Minu, Automatic Emotion Recognition and Classification, Procedia Engineering, Volume 38, 2012, Pages 21-26, ISSN 1877-7058

[2] Thapanee Seehapoch and Sartra Wongthanavasu, “Speech emotion recognition using support vector machines,” in Knowledge and Smart Technology (KST), 2013 5th International Conference on. IEEE, 2013, pp. 86–91.

[3] Bjorn Schuller, Gerhard Rigoll, and Manfred Lang, ¨ “Hidden markov model-based speech emotion recognition,” in Multimedia and Expo, 2003. ICME’03. Proceedings. 2003 International Conference on. IEEE, 2003, vol. 1, pp. I–401.

[4] Kun Han, Dong Yu, and Ivan Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[5] Dario Bertero and Pascale Fung, “A first look into a convolutional neural network for speech emotion detection,” in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017, pp. 5115–5119.

[6] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik, “Speech emotion recognition from spectrograms with deep convolutional neural network,” in Platform Technology and Service (PlatCon), 2017 International Conference on. IEEE, 2017, pp. 1–5.

[7] John Gideon, Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost, “Progressive neural networks for transfer learning in emotion recognition,” Proc. Interspeech 2017, pp. 1098–1102, 2017.

[8] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, “Spoken emotion recognition using hierarchical classifiers,” Comput. Speech Lang., vol. 25, no. 3, pp. 556–570, Jul. 2011.

[9] S. Wu, T. H. Falk, and W.-Y. Chan, “Automatic speech emotion recognition using modulation spectral features,” Speech Commun., vol. 53, no. 5, pp. 768–785, May 2011.

[10] . L. Nwe, S. W. Foo, and L. C. De Silva, “Speech emotion recognition using hidden Markov models,” Speech Commun., vol. 41, no. 4, pp. 603–623, Nov. 2003.

[11] M. M. H. El Ayadi, M. S. Kamel, and F. Karray, “Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models,” in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP ’07, 2007, vol. 4, pp. IV–957–IV–960.

[12] Wadhwa, M., A. Gupta, and P. K. Pandey. "Speech emotion recognition (ser) through machine learning." (2020). From https://www.analyticsinsight.net/speech-emotion-recognition-ser-through-machine-learning.

[13] S. Kavitha, N. Sanjana, K. Yogajeeva and S. Sathyavathi, "Speech Emotion Recognition Using Different Activation Function," 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), 2021, pp. 1-5, doi: 10.1109/ICAECA52838.2021.9675789.

[14] P. Tzirakis, A. Nguyen, S. Zafeiriou and B. W. Schuller, "Speech Emotion Recognition Using Semantic Information," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6279-6283, doi: 10.1109/ICASSP39728.2021.9414866.

[15] P. A. Babu, V. Siva Nagaraju and R. R. Vallabhuni, "Speech Emotion Recognition System With Librosa," 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), 2021, pp. 421-424, doi: 10.1109/CSNT51715.2021.9509714.

[16] M. Sakurai and T. Kosaka, "Emotion Recognition Combining Acoustic and Linguistic Features Based on Speech Recognition Results," 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), 2021, pp. 824-827, doi: 10.1109/GCCE53005.2021.9621810.

[17] Parimala, M., Swarna Priya, R. M., Praveen Kumar Reddy, M., Lal Chowdhary, C., Kumar Poluru, R., & Khan, S. (2021). Spatiotemporal-based sentiment analysis on tweets for risk assessment of event using deep learning approach. Software: Practice and Experience, 51(3), 550-570.

[18] Rodrigues, A. P., Fernandes, R., Shetty, A., Lakshmanna, K., & Shafi, R. M. (2022). Real-Time Twitter Spam Detection and Sentiment Analysis using Machine Learning and Deep Learning Techniques. Computational Intelligence and Neuroscience, 2022.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in neural information processing systems, 2014, pp. 2672–2680.

[20] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” arXiv preprint arXiv:1511.06434, 2015.

[21] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to image translation using cycle-consistent adversarial networks,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[22] A. Antoniou, A. J. Storkey, and H. A. Edwards, “Data augmentation generative adversarial networks,” CoRR, vol. abs/1711.04340, 2018.

[23] X. Zhu, Y. Liu, Z. Qin, and J. Li, “Data augmentation in emotion classification using generative adversarial networks,” arXiv preprint arXiv:1711.00648, 2017.

[24] [24] S. Sahu, R. Gupta, and C. Espy-Wilson, “On enhancing speech emotion recognition using generative adversarial networks,” Interspeech 2018, September 2018.

[25] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi, “Bagan: Data augmentation with balancing gan,” arXiv preprint arXiv:1803.09655, 2018.

[26] Y. Li, C. Baidoo, T. Cai and G. A. Kusi, "Speech Emotion Recognition Using 1D CNN with No Attention," 2019 23rd International Computer Science and Engineering Conference (ICSEC), 2019, pp. 351-356, doi: 10.1109/ICSEC47112.2019.8974716.

[27] S. Saleem, A. Dilawari, M. U. Ghani Khan and M. Husnain, "Voice Conversion and Spoofed Voice Detection from Parallel English and Urdu Corpus using Cyclic GANs," 2019 International Conference on Robotics and Automation in Industry (ICRAI), 2019, pp. 1-6, doi: 10.1109/ICRAI47710.2019.8967385.