Speech Emotion Recognition Using ANN on MFCC Features

Harshit Dolka
Computer Science and Engineering
Karunya Institute of Technology and Sciences
Tamil Nadu, India
harshkr768@gmail.com

Arul Xavier V M
Computer Science and Engineering
Karunya Institute of Technology and Sciences
Tamil Nadu, India
arul_vmax@karunya.edu

Sujitha Juliet
Computer Science and Engineering
Karunya Institute of Technology and Sciences
Tamil Nadu, India
sujitha@karunya.edu

2021 3rd International Conference on Signal Processing and Communication (ICPSC) | 978-1-6654-2864-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICSPC51351.2021.9451810

Abstract—**Speech Emotion Recognition (SER) is one of the active research topics in Human-Computer Interaction. This paper focuses on training an ANN Model for SER using Mel Frequency Cepstral Coefficients (MFCCs) feature extraction and training it on selected audio datasets to compare the performance. The model can classify audio files based on a total of eight emotional states: happy, sad, angry, surprise, disgust, calm and neutral, although the number of emotions varies in selected datasets. The proposed model gives an average accuracy of 99.52% on the TESS data set, 88.72% on the RAVDESS data set, 71.69% on the CREMA data set, and 86.80% on the SAVEE data set.**

Keywords—Speech Emotion Recognition, Artificial Neural Networks, Multi-layer Perceptron, Audio datasets, Deep Learning, MFCC Features

I. INTRODUCTION

As human beings, speech is the most widely used mode of communication among us. Speech contains a lot of behavioural cues and information about our mentality and emotions, apart from just the words that we speak. When we speak, we use language as the medium and we make phonetic combinations of different vowels and consonant sounds to form words. Now, the purpose of speaking also includes conveying our feelings or emotions, and voluntarily or involuntarily, humans tend to convey emotions through their speech itself. Based on their emotions, their tone, frequency, loudness, and break in-between the words are affected, which is easily recognizable to human ears, but it is not so easy for machines like computers to recognize the same. Therefore, emotion recognition based on speech becomes a challenging task, and thus, to make machines capable of recognizing human emotions through their speech pattern, we use a concept called Speech Emotion Recognition (SER).

SER is a technique used by machines to detect human emotions based on their speech. When deployed efficiently, SER has many real-world applications such as personalized music/movie player, business marketing, and suicide prevention.

In this paper, an ANN with ReLU model using MFCC feature extraction has been proposed, which has been trained on 4 selected datasets, and the performances of the model on each of the datasets have been compared. The reasons for going with ANN over CNN as the algorithm of choice are mentioned below:

1) CNN algorithms require image datasets or the conversion of the audio files to image datasets to perform in their optimal capacity. But, storing image files requires much more memory than audio files, and therefore it is a memory-consuming job.

2) Traditional CNN models require more time to train on the image datasets than it takes for the ANN models to get trained on the features of the audio datasets, therefore, the training of CNN models is time-consuming.

For ease of use, this paper has been divided into 5 sections. Section 1 gives the introduction to the paper topic. In section 2, a literature survey on previous works related to speech emotion recognition using various methodologies has been covered. The proposed model has been explained in section 3. In section 4, experimental testing of the model has been performed on various selected datasets and the performance parameters have been provided for the same. The paper then gets concluded in section 5, discussing a few observations and the future scope of the model. It has been then followed by references.

II. LITERATURE SURVEY

As SER is a very active field of research across the world of technology, there are many innovations and research work that has been done in this domain in the past.

In Yi-Lin Lin et al. [1], a Hidden Markov Model (HMM) based SER was proposed which used five distinct features in contrast to popularly used MFCC features set for classification of five emotions on Danish Emotional Speech Database (DES) and yielded an accuracy of 99.5% for the gender-independent case. Support Vector Machine (SVM) based SER was also tested on the same databases using Mel Energy Spectrum Dynamic Coefficients (MEDC) feature extraction and it yielded an accuracy of 88.9% for the gender-independent case.

More advanced methods were proposed by several researchers later on, including Abdel-Hamid et al. [2], which proposed Convolutional Neural Networks (CNN) and Zheng et al. [3], which used Deep Convolutional Neural Networks (DCNN) for SER, and achieved 40.02% accuracy rate, since the CNN methodology was in its initial stages.

But further developments were made in using the Image-Based CNN algorithm for speech emotion recognition and in W Lin et al. [4], a precision rate of 87.74% was achieved using CNN for Berlin Emotional Database. CNN proved to be a promising method when dealing with image classification and image feature distinction. Since the speech segments were first converted into their respective spectrograms, it was then easy to apply the CNN algorithm to them to build a neural model.

Another popular field of neural networks called Artificial Neural Network (ANN) has also been used for SER. ANN uses a feed-forward neural network that builds on a given input weight and then processes it to find the least error value. Several layers are added to the model to get the single output. Sequential ANN is useful when the model is of a single input-single output type.

It was observed in Zeidan et al., [5] that the results obtained over varying coefficient numbers by using ANN were slightly less than Support Vector Machine (SVM), which is yet another popularly used classifier method for SER. Jain et al. [6] proposed an SER model based on SVM, using MFCC, MEDC, and energy features which achieved a recognition rate of 91.30% and 95.08% for EmoDB and SJTU Chinese Databases respectively. SVM is used in supervised learning, and it is effective in both classification and regression problems.

Another descendant of ANN, Recurrent Neural Network, also began to gain acceptance in the SER world. RNNs use something called LSTM (Long Short-Term Memory) to keep track of their input, making it the first of its kind to do so. It is this distinctive feature that makes it very suitable for problems that have sequential data. In 2017, Mirsamadi et al. [7] proposed an RNN based SER model in which they used weighted-pooling with local attention, which was better than compared to traditional SVM solution by +5.7%.

These works have been accomplished by using several algorithms such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Artificial Neural Networks (ANN), MLP Classifier, Support Vector Machine (SVM), and Hidden Markov Model (HMM) to name a few. In this paper, an SER model has been trained on different selected datasets, using multi-layer perceptron or ANN.

However, the main reason why these models cannot be compared with each other is because of two reasons:

1) Different models build on different feature sets of audio files, and this affects their performance. While most of the models make use of the global MFCC feature extraction technique, others also use LLDs, MEDC, and energy features, or a combination of them to yield better results.

2) There are no universal audio datasets available in the present, and therefore different models use different datasets to train their Deep Learning Models, which further makes it harder to compare the performance of different algorithms.

In this paper, an ANN with ReLU model using MFCC features on selected datasets has been proposed. A comparison of the performance parameters of the trained model on 4 selected datasets has also been provided.

III. Proposed Work

A. Deep Learning

To make computers understand and imitate human actions and behavior, Artificial Intelligence (AI) is the field often looked forward to, and the process of making the machines learn what we do can be termed as Machine Learning (ML). ML comprises many such algorithms and techniques that can be trained on a given set of data, and Deep Learning (DL) is a part of it.

In DL, multi-layered neural networks are trained using huge amounts of data, often stored in datasets. Now, the type of learning for a DL model can be of three types - unsupervised, supervised, and semi-supervised learning based on the type of datasets provided for training. DL has many different algorithms available for different types of problems. Some of them are CNN (Convolutional Neural Network), ANN (Artificial Neural Network), and RNN (Recurrent Neural Networks). In this paper, the model is using the ANN algorithm to predict emotions based on speech input.

B. Artificial Neural Networks (ANN)

ANN is a DL algorithm that contains connected units called artificial neurons, which take data from the input layer, and transfer data from one layer to another (if there is more than one layer), and determine the output. It works similarly as our human brain does, but not that complexly. A single-layered neural network is called a perceptron, and ANN is a multi-layered perceptron.

C. Datasets Used

For this paper, 4 datasets have been used for individual training and testing of the proposed ANN model. The selection of the datasets was based on keeping the English language as the medium and also to bring out not-so-common datasets currently existing, into the limelight by using them for the training of the ANN model. The details of the selected datasets are given below:

1) RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [8]: This data set contains both audio and video files of 24 professional actors (12 male, 12 female), speaking two sentences in English language. But since we are working on speech input, only audio files have been taken into consideration. This data set contains 1440 audio-only files classifies into 8 emotions. For this paper, only the 6 most common emotions out of them have been considered, namely - happy, sad, angry, fear, surprise, and calm to achieve better results.

2) TESS (Toronto Emotional Speech Set) [9]: This data set contains a total of 2800 audio files, which are recorded in the voices of two actresses (aged 26 and 64).

3) CREMA-D (Crowd-Sourced Emotional Multimodal Actors Data Set) [10]: This is the largest of all the databases selected for this paper. It contains 7442 audio files from 91 actors (48 male and 43 female) and is classified into six emotions namely: happy, sad, anger, disgust, fear, and neutral.

4) SAVEE (Surrey Audio-Visual Expressed Emotion) [11]: This data set contains 480 audio files recorded in the voices of 4 actors aged between 27 to 31 years. There are a total of 120 utterances per speaker.

D. Feature Extraction

SER is a process involving the conversion of speech input into digital signals and then processing it to extract related features suitable for training the model.

Some of the important extracted features in the proposed system are given below:

1) Zero-Crossing Rate (ZCR):

The Zero-Crossing Rate (ZCR) of an audio frame is the rate of sign-changes of the signal during the time frame. In other words, it is the number of times the signal changes its value, from positive to negative and also from negative to positive, divided by the length of the frame. The ZCR is defined by the following equation:

$$Z(i) = \frac{1}{2W_L} \sum_{n=1}^{w_l} | \mathrm{sgn}[x_i(n)] - \mathrm{sgn}[x_i(n-1)] | \qquad (1)$$

where sgn(·) is the sign function, i.e.

$$\mathrm{sgn}[x_i(n)] = \begin{cases} 1, & x_i(n) \geq 0, \\ -1, & x_i(n) < 0 \end{cases} \qquad (2)$$

2) Mel-Frequency Cepstral Coefficients (MFCCs): MFCC features extraction technique is used to convert the conventional frequency to Mel scale.

The formula used to calculate the mels for any frequency is given below [12]:

$$\mathrm{mel}(f) = 2595x \log_{10}(1 + f/700) \qquad (3)$$

where mel(f) is the frequency (mels) and f is the frequency (Hz).

The MFCCs are calculated using this equation [12, 13]:

$$\hat{C}_n = \sum_{n=1}^{k} (\log \hat{S}_k) \cos[n(k - \tfrac{1}{2}) \tfrac{\pi}{k}] \qquad (4)$$

where k is the number of mel cepstrum coefficients,   k is the output of filter bank and   n is the final MFCC coefficients.

3) Chroma STFT: It is used to create a chromogram from a wave spectrogram.

4) Root-Mean-Square Value (RMS): RMS computes the RMS value for each frame.

5) Mel-scaled Spectrogram: Mel-Scaled Spectrogram is a spectrogram where frequencies are converted into mel-scale frequencies.

E. Proposed SER Model:

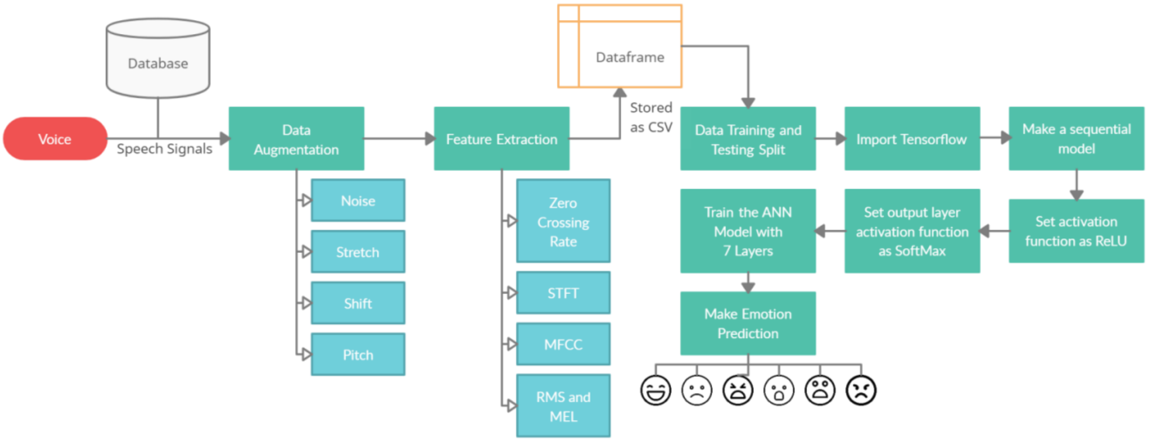

Fig. 1. Proposed SER System

In the proposed model as shown in figure 1, the datasets are loaded from the database and are fed further into the system in the form of speech signals.

Next, data augmentation techniques are used to make the datasets more similar to real-world audio input, by using the following steps:

1) Noise Injection: It adds some random values into the data by using NumPy. Noise injection is a very good augmentation technique because of which it can be assured that the trained model is not overfitted.

2) Shifting: It just shifts the audio to left/right with a random second. If shifting audio to left (fast forward) with x seconds, the first x seconds will mark as 0.

3) Stretching: It stretches times series by a fixed rate.

4) Pitch Change: It changes pitch randomly.

After the data has been augmented, the Librosa library is used to extract selected audio features, and store them in a data frame in the CSV format.

As the next step, the data set is divided into the testing part and the training part. In the proposed model, 90% of data has been taken for training purposes and 10% of data has been taken for training purposes. After that, the ANN model is trained. After the model is trained, the model is then made to predict emotions from given audio files, and the performance evaluation is given in Table I.

IV. Experimental Results and Analysis

A. ANN Model

The proposed model uses supervised learning for the datasets since the data points are labeled. As the first step, Tensorflow has been imported to implement the sequential model for the research. The proposed model is a feed-forward sequential ANN model, which means that the information can only travel in the forward direction, from the input layer to the output layer, going through one or many hidden layers.

1) Sequential ANN Model: Sequential models are one of the simplest artificial neural networks and are most efficient when the model has a single input and single output per unit time, which also holds in this case.

2) Activation Function Used: The activation function that is used for input and hidden layers is 'ReLU'. ReLU or Rectified Linear Unit function is one of the most common functions used in ANNs. Its functionality includes returning 0 if it receives any negative value as input, and returning the same value which it has received in case of the value being positive. Therefore, it can be summarized as

$$f(x) = \max(0, x) \tag{5}$$

ReLU helps the model to learn faster and perform better, and it also helps solve the vanishing gradient problem. As its computation is fairly simple, it is suitable for the model. For the output layer, softmax activation function has been used.

3) Softmax Activation Function: The Softmax function is commonly used in both regression and classification models. Since the model is of the multi-class classification type, it is suitable for the model. Softmax is defined by the function:

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{6}$$

Then the model is trained by passing 162 as input weight and output unit as 1024 in the input layer. The model has 7 layers, and finally gives an output possibility of 6 or 7 classes depending on which data set has been selected to train it.

4) Validation Method Used: For this model, the cross-validation method has been used to validate the model and obtain validation accuracy and loss. In the cross-validation method, a small part of the training data set is kept hidden from the model, and the model is later made to run through the unseen part to predict the labels. In this model, 10% of the data set has been used as the validation part and the rest 90% has been used for training the model.

5) Loss Function Used: For this model, categorical cross-entropy has been used as the loss function, as it is the combination of softmax function and cross-entropy loss and also because it is commonly used for multi-class classification models.

The categorical cross-entropy loss function calculates the loss of an example by computing the following formula [14]:

$$\text{Loss} = -\sum_{i=1}^{\text{output size}} y_i \cdot \log \hat{y}_i \tag{7}$$

where $\hat{y}_i$ is the i-th scalar value in the model output, y_i is the corresponding target value, and the output size is the number of scalar values in the model output.

B. Performance Evaluation

The proposed model was trained on the selected datasets, and the results are compared in Table I.

TABLE I. ACCURACY COMPARISON OF THE PROPOSED MODEL ON SELECTED DATASETS

Type of Model	Datasets Used For Training			
	RAVDESS	*TESS*	*CREMA-D*	*SAVEE*
ANN+ReLU (MFCC)	88.72%	**99.52%**	71.69%	86.80%

It is evident that the proposed ANN+ReLU model worked almost perfectly on Toronto Emotional Speech Set (TESS) and gave an accuracy of 99.52%. The emotions that were recognized the best by the model on this data set were anger, fear, and neutral with 100% precision. The proposed model also gave a higher accuracy on the RAVDESS data set having 6 emotions than previous state-of-the-art models proposed in earlier works. Table II compares the performances of different models on the RAVDESS data set with that of the proposed model.

TABLE II. ACCURACY COMPARISON OF DIFFERENT MODELS ON THE RAVDESS DATA SET

SER Model	Accuracy
Gao et al. [15]	79.4%
Issa et al. [16]	71.61%
Sajjad et al. [17]	77.02%
Shegokar et al. [18]	60.01%
Bhavan et al. [19]	75.69%
Proposed Model (6 emotions)	**88.72%**

V. Conclusion

For this paper, we worked on training a multi-layered perceptron ANN with ReLU model which can detect speech emotions by taking the audio file as its input. In this paper, the proposed model was trained on four selected datasets and the performance parameters of the model on selected datasets were compared with each other. It was observed that the proposed model trained using MFCC feature extraction is the most efficient on the TESS dataset giving 99.52% accuracy, while it also gave high accuracies on the RAVDESS and the SAVEE datasets. The model did not prove to be so efficient on the CREMA dataset. The model performance on the RAVDESS data set was also compared with other state-of-the-art methods from other works, and it was seen that the proposed model achieved higher accuracy than them.

As the further scope of this paper, other popular audio datasets can also be used to train the proposed model individually and their performances can be compared. Also, a hybrid ANN-CNN model can be created where the CNN algorithm can be applied to read and process facial expressions to predict emotions, and the ANN algorithm can be used to predict emotions based on speech features.

References

[1] Lin, Y.L. and Wei, G., 2005, August. Speech emotion recognition based on HMM and SVM. In 2005 international conference on machine learning and cybernetics (Vol. 8, pp. 4898-4901). IEEE.

[2] Abdel-Hamid, O., Mohamed, A.R., Jiang, H., Deng, L., Penn, G. and Yu, D., 2014. Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing, 22(10), pp.1533-1545.

[3] Zheng, W.Q., Yu, J.S. and Zou, Y.X., 2015, September. An experimental study of speech emotion recognition based on deep convolutional neural networks. In 2015 international conference on affective computing and intelligent interaction (ACII) (pp. 827-831). IEEE.

[4] Lim, W., Jang, D. and Lee, T., 2016, December. Speech emotion recognition using convolutional and recurrent neural networks. In 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA) (pp. 1-4). IEEE.

[5] Zaidan, N.A. and Salam, M.S., 2016. MFCC global features selection in improving speech emotion recognition rate. In Advances in machine learning and signal processing (pp. 141-153). Springer, Cham.

[6] Jain, M., Narayan, S., Balaji, P., Bhowmick, A., and Muthu, R.K., 2020. Speech emotion recognition using support vector machine. arXiv preprint arXiv:2002.07590.

[7] Mirsamadi, S., Barsoum, E. and Zhang, C., 2017, March. Automatic speech emotion recognition using recurrent neural networks with local attention. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2227-2231). IEEE.

[8] Steven R. Livingstone and Frank A. Russo, “RAVDESS Emotional speech audio.” Kaggle, 2019, doi: 10.34740/KAGGLE/DSV/256618.

[9] https://www.kaggle.com/ejlok1/toronto-emotional-speech-set-tess

[10] https://www.kaggle.com/ejlok1/cremad

[11] https://www.kaggle.com/barelydedicated/savee-database\

[12] Chakroborty, S., Roy, A. and Saha, G., 2006, December. Fusion of a complementary feature set with MFCC for improved closed set text-independent speaker identification. In 2006 IEEE International Conference on Industrial Technology (pp. 387-390). IEEE.

[13] Alim, S.A. and Rashid, N.K.A., 2018. Some commonly used speech feature extraction algorithms (pp. 2-19). IntechOpen.

[14] https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/loss-functions/categorical-crossentropy

[15] Gao, Y., Li, B., Wang, N. and Zhu, T., 2017, November. Speech emotion recognition using local and global features. In *International Conference on Brain Informatics* (pp. 3-13). Springer, Cham.

[16] Issa, D., Demirci, M.F. and Yazici, A., 2020. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, *59*, p.101894.

[17] Sajjad, M. and Kwon, S., 2020. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access*, *8*, pp.79861-79875.

[18] Shegokar, P. and Sircar, P., 2016, December. Continuous wavelet transform based speech emotion recognition. In *2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS)* (pp. 1-8). IEEE.

[19] Bhavan, A., Chauhan, P. and Shah, R.R., 2019. Bagged support vector machines for emotion recognition from speech. *Knowledge-Based Systems*, *184*, p.104886.