International Journal of Recent Trends in Technology and Engineering
ISSN [O]:2832-4277 Volume-II Issue-II April - June-2023

Identification of Emotions Through Voice

Naveenkumar R[1], Privietha P [2]

[1]Student, Department of Computer Applications, Hindusthan College of Engineering and Technology, Coimbatore, naveenkumar11301@gmail.com

[2]Assistant Professor, Department of Computer Applications, Hindusthan College of Engineering and Technology, Coimbatore, priviethaprabhakar@gmail.com

Abstract- ***The article describes a research project in the field of Human-Computer Interaction, which focuses on the problem of automatic emotion recognition from speech. The aim of the project is to create a system that can recognize emotional states in the same way as humans to make the interaction between humans and digital machines more natural. The researchers used a modified recurrent neural network (RNN) architecture with long short-term memory (LSTM) to extract multiple temporal features from the audio signal. The experiment was done using an open dataset that includes eight different emotions: neutral, calm, happy, sad, angry, scared, disgust, and surprised. The accuracy of the model was tested using 40% of the dataset. The dataset used in the experiment is the TESS dataset, which includes 2800 audio data samples of two females speaking 200 target words with seven different emotions. The research used Python language and the Keras package, which is an open-source neural network library written in Python. The research found that the proposed deep learning - RNN architecture showed an outstanding performance on various problems.***

Keywords: ***Speech Emotion Recognition, Deep Learning, Audio, RNN, LSTM.***

1. INTRODUCTION

Speech Emotion Recognition (SER) is a rapidly growing field with significant implications for various applications in today's digital era. As technology advances, machines can interpret and recognize emotions from human speech signals, leading to advancements in fields such as automatic translation systems, machine-to-human interaction, and speech synthesis.

Emotion recognition is the study of identifying the six universal expressions (anger, joy, fear, happiness, sadness, and surprise) represented in figure1 using computer science techniques. These emotions reflect one's state of mind, and recognizing them is crucial for developing machines that can simulate human-like interactions.

In addition to recognizing emotions, valence and arousal are essential factors in identifying one's state of mind. Sentiment analysis is used to understand a person's opinion and attitude towards a particular topic or at a specific time using various computational approaches.

Affective computing or artificial emotional intelligence is the study of human emotions, their interpretation, processing, and adaptation by machines. Machines can recognize human emotional states from various sources, such as facial expressions, body movements, speech, text writing, brain, or heart signals. Machine learning techniques are used to extract required features or patterns from the collected data, which help machines simulate human-like interactions.

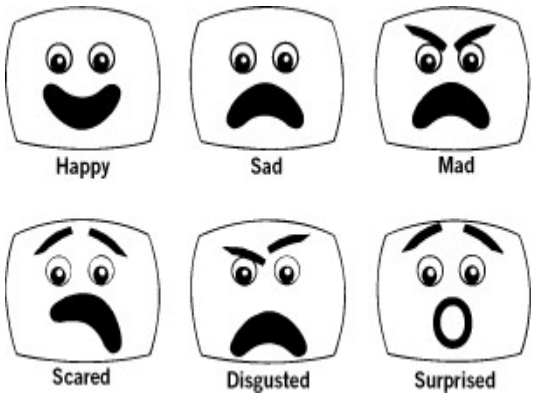

Figure1: Emotions

RNN and LSTM:

Both RNNs and LSTMs are types of neural networks that are used for processing sequential data. However, while RNNs are capable of recognizing the sequential characteristics of data, they suffer from the vanishing gradient problem when the sequence becomes too long. This problem arises when the gradient of the loss function becomes so small that it effectively stops the network from learning.

LSTMs are a type of RNN that solve the vanishing gradient problem by using a memory cell that can maintain its state over time. The LSTM architecture consists of several gates that

control the flow of information into and out of the memory cell, which makes it capable of learning long-term dependencies in sequential data.

In summary, RNNs are capable of processing sequential data, but they may suffer from the vanishing gradient problem when the sequence becomes too long. LSTMs are a type of RNN that solve this problem by using a memory cell and gates to control the flow of information. LSTMs are therefore capable of learning long-term dependencies in sequential data and are widely used in applications such as speech recognition and machine translation.

2. RELATED WORK

Deep learning has revolutionized many areas of artificial intelligence and has enabled the development of powerful applications such as speech recognition, image recognition, and natural language processing [1]. Recurrent neural networks, including the LSTM architecture, are a type of deep learning model that can capture sequential dependencies in data, making them well suited for time-series data, speech recognition, and natural language processing tasks [2]. LSTM networks have been shown to perform exceptionally well on a wide range of problems due to their ability to handle long-term dependencies in the input data [3].

There is a need for accurate emotion recognition in various fields, including human-computer interaction [4]. The proposed models achieved high accuracy in recognizing emotions from speech signals, ranging from 91% to 97.05% [5]. The use of image classification methods for emotion recognition and achieved an average accuracy of 94.05% [6]. Many authors emphasize that accurate emotion recognition is crucial for effective communication and human interaction with machines [7].

Speech emotion recognition is a significant research topic in the field of artificial intelligence and human-computer interaction. Machine learning and deep learning techniques are commonly used to develop emotion recognition models, which aim to identify emotional or physical states of a person from their speech signals. Various features such as prosody, Mel-frequency cepstrum coefficients (MFCC), modulation spectral (MS), and Teag-energy-operator (TEO) have been used for feature extraction. Different classifiers are also used to train these models. The average accuracy achieved by the proposed classifiers ranges from 90% to 97.05%. It can be seen that the accuracy of these models is relatively high, indicating that they have the potential to be used in practical applications.

3. METHODOLOGY

In this research, the investigator used python language to develop the basic deep learning – RNN architecture. A google engineer has developed keras an open source neural network library written in Python language. The default library packages in keras are imported to simplify the work of the investigator. Figure 2 represents the methodology of the model extraction.

The modified RNN is designed with the help of keras package to improve the accuracy. LSTM package is directly imported and utilized along with dense and dropout layers. Activation function uses “Relu” and “Softmax” for statistical calculations. Relu is used for activation of nodes and softmax is used for probability calculation in the last activation layer of the RNN.

The total dataset is divided into 2 phases 60% for training and 40% for testing. Using 60% of training, the model is framed in deep learning and the 40% is used to check out the accuracy of the predicted model.

Epoch size is set to 100 and the batch-size is set as 512. Shuffling is set as true for random picking of dataset. For each epoch training accuracy, training loss, validation accuracy, validation loss to calculated and tabulated. Using plotting result the validation loss and validation accuracy is calculated.

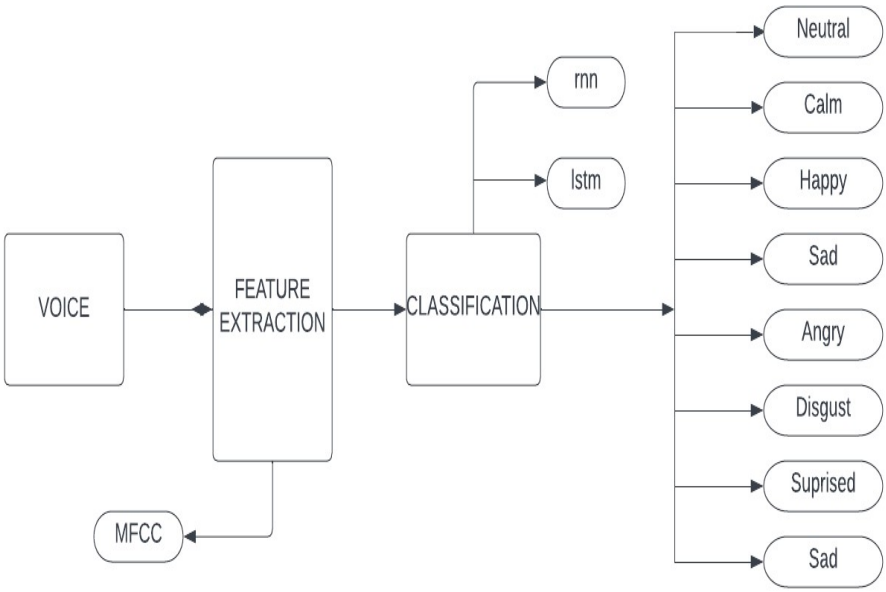

Figure 2: Methodology

4. DATASET:

TESS dataset were modelled on the North Western University Auditory Test No. 6 by Kate Dupuis, M. Kathleen Pechora - Fuller in University of Toronto, Psychology Department, 2010. This collection is published under Creative Commons license Attribution-No Commercial – No Derivatives 4.0 International.

Two females were recruited from the Toronto area. Both females speak English as their first language, are university educated, and have musical training. Audiometric testing indicated that both females have thresholds within the normal range.

A set of 200 target words were spoken by two females are made by the set portraying the seven emotions. There are 2800 data's in total.

A dataset for training of seven cardinal emotions classification in audio speech recognition. Figure 3 represents the difference of sad and happy emotions

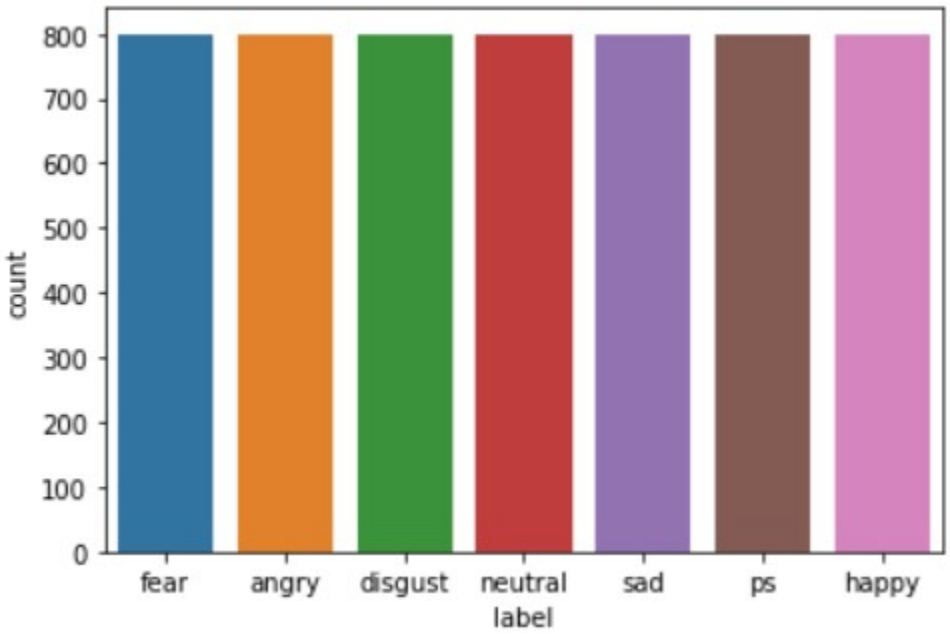

Figure 3: Difference of Sad and Happy Emotions

5. NETWORK ARCHITECTURE

By using the library packages in keras a new deep learning, Recurrent neural network architecture is designed with various layers like, activation, soft max, dropout and dense in this research work.

The activation function decides whether a neuron should be activated by calculating the weighted sum and further adding bios to it. The purpose of the activation function is to introduce non-linearity into the result of a neuron. When comparing with a neuron-based model that is in our brains, the activation function is at the end deciding what is to be fired to the next neuron.

Soft max assigns decimal probabilities to each class in a multi-class problem. Those decimal probabilities must add up to 1.0. This additional constraint helps training converge more quickly than it otherwise would. Soft max is implemented through a neural network layer just before the output layer.

Dense Layer is used to analyse image based on output from convolutional layers. Every Layer in Neural Network contains neurons, it compute the weighted average of its input and this weighted average is moved through a nonlinear function, called as an activation function.

The dropout used to dropping out the nodes (input and hidden layer) in a neural network. Every forward and backwards connections with a dropped node are temporarily removed, so creating a new network architecture out of the parent network.

Activation (ReLU), dropout (0.5), dense (4), activation (softmax) layers are executed again to get better accuracy and performance. The softmax function is often called in the final stage because it changes the output to 0 and 1 with 1 as each total probability sum. There are no parameters required for the softmax function.

```
from keras.models import Sequential
from keras.layers import Dense, LSTM, Dropout

model = Sequential([
    LSTM(123, return_sequences=False, input_shape=(40,1)),
    Dense(64, activation='relu'),
    Dropout(0.2),
    Dense(32, activation='relu'),
    Dropout(0.2),
    Dense(7, activation='softmax')
])

model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
model.summary()
```

Figure 4: Network Architecture using Python

6. TRAINING AND TESTING:

Figure 4, 5 and 6 represents the coding part in Jupiter notebook for model preparation. Training a model simply means learning good values for every weights from labelled examples. In supervised learning, a machine-learning algorithm builds a model by examining lot of examples and attempting to find a model that reduce loss this process is called empirical risk minimization.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 123)               61500
_________________________________________________________________
dense (Dense)                (None, 64)                7936
_________________________________________________________________
dropout (Dropout)            (None, 64)                0
_________________________________________________________________
dense_1 (Dense)              (None, 32)                2080
_________________________________________________________________
dropout_1 (Dropout)          (None, 32)                0
_________________________________________________________________
dense_2 (Dense)              (None, 7)                 231
=================================================================
Total params: 71,747
Trainable params: 71,747
Non-trainable params: 0
_________________________________________________________________
```

Figure 5: Layer Design

7. RESULTS:

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Find a reliable source with complete, correct, and verifiable information. Accuracy refers to the correctness, truthfulness, and overall excellence and quality of the information.

```
epochs = list(range(100))
acc = history.history['accuracy']
val_acc = history.history['val_accuracy']

plt.plot(epochs, acc, label='train accuracy')
plt.plot(epochs, val_acc, label='val accuracy')
plt.xlabel('epochs')
plt.ylabel('accuracy')
plt.legend()
plt.show()
```

Figure 6: Python Code for Calculating Accuracy

Figure 7 and 8 displays the pictorial representation of validation loss and accuracy. The graph shows a stable growth in accuracy for each epoch.

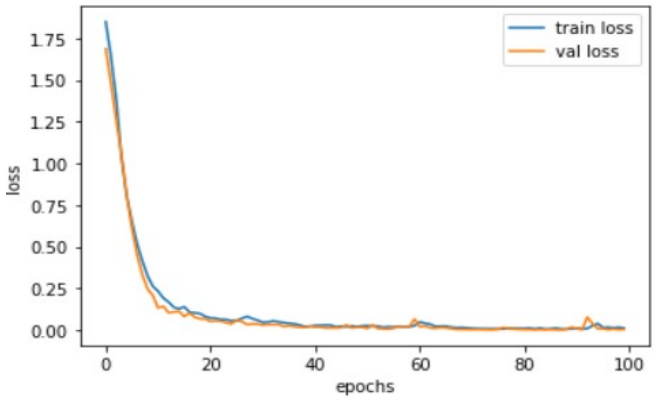

Figure 7: Pictorial Representation of Training Loss and Validating Loss

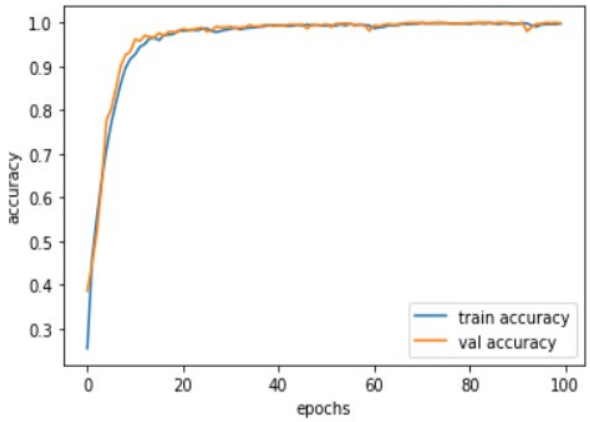

Figure 8: Pictorial Representation of Training Accuracy and Validating Accuracy

8. CONCLUSION

This paper provides survey of many research papers based upon emotion recognition, sentimental analysis, applications of emotion recognition systems and supervised & unsupervised machine learning algorithms required for automatic emotion recognition. It has been revealed that self-learning algorithm Recurrent neural networks produces good results for naturalistic databases, also best fitted to reduce data over fitting and data imbalance. Emotion recognition provides benefits to many institutions and aspects of life. It is useful and important for security and healthcare purposes. Also, it is crucial for easy and simple detection of human feelings at a specific moment without actually asking them.

REFERENCES

1. Bettadapurai Vinay, "Face action recognition and analysis: the state of art", *preprint arXiv:1203.6722*, 2013.
2. Matthew S Ratliff and Eric Patterson, "Emotion recognition using facial expressions with active appearance models".
3. Samal and P.A. Iyyangar, "Automatic Recognition and Analysis of Human Face and Facial Expressions: A Survey", Pattern Recognition, vol. 25.
4. K.R. Scherer and H. Ellgring, "Multimodel Expression of Emotion: Affect Programsn Emotion", vol. 7.
5. Emerich Simina, Eugen Lupu and Anca Apatean, "Emotions recognition by speech and facial expressions analysis", *Proceedings of the 17th European Signal Processing Conference (EUSIPCO'09)*, pp. 1617-1621, 2009.
6. Dino, H.I. and Abdulrazaq, M.B. (2019) Facial Expression Classification Based on SVM, and MLP Classifiers. 2019 International Conference on Advanced Engineering.
7. T. Mohana Priya, Dr. M. Punithavalli & Dr. R. Rajesh Kanna “Global Journal of Computer Science and Technology” C Software & Data Engineering, Volume 20, Issue 2, No. 2020, pp 12-20
8. Krizhevsky, A., Sutskever, I., (2013). Image net classification from deep convolutional neural network. Adv. Neural Inform. Process. Syst. 2012, 1097–1105.
9. Krothapally, S. R., and Kolagudi, S.D. (2013). Emotion Recognition Using Speech quality (New York, Springer-Verlag) 10.1007/978-1-4614-5143-3
10. Lech, M., Stolar, M., Bolia, R., (2017). Amplitude-frequency survey of emotional speech by use of transfer learning and classification of spectrogram images. Adv. Eng. Technol. Eng. Syst. J. 3, 363–371.