International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 1300-1306 www.ijaem.net ISSN: 2395-5252

Speech Emotion Recognition: An Art of DetectingEmotions through Machines

[1]Shreya Taware , [2]Sarang Kulkarni , [3]Vaishnavi Taware, [4]Pooja Tilekar

[1,2,3,4]*B.E Computer Engineering, Department of Computer Engineering*
[1,2,3,4]*Savitribai Phule Pune University,Pune.*
[1,2,3,4] *Vidya Pratisthan's Kamalnayan Bajaj Institute of Engineering and Technology, Maharashtra,India*

Submitted: 01-06-2021 Revised: 14-06-2021 Accepted: 16-06-2021

ABSTRACT: In the process of training machines to think like human brains, understanding human emotions through his/her speech plays a vital role in Human-Computer Interaction (HCI). This process involves various steps including data collection, pre-processing, feature extraction, feature selection, normalization, building a classification model with the best accuracy, and thus detecting the precise emotion of the speaker. This paper covers the discussion over the work done in this field of Speech Emotion Recognition (SER) along with the failed trials and models designed to get the perfect accuracy.

Keywords : Speech, Emotion, Recognition, Convolutional, Neural, Network, Mel, frequency, Coefficients.

I. INTRODUCTION:

Speech has always been the most effective way of conversation. We, as humans are trained to identify emotions through a speech from a very small age but making machines understand them as a part of Human-Computer Interaction is a bit difficult. For this, we need to perform a very lengthy process of training the model to give us the most accurate results. Speech emotion recognition is a very crucial field in HCI. We can quote SER as:

Speech Emotion Recognition System is the collection of methods that process the sound signals to extract the features in the signal, classify those features with the deep learning networks to detect the emotion in a signal.[1]

Let's understand Speech Emotion Recognition in detail. When we hear a person talk our brain processes his voice and collects various features from his voice. These features are further processed to identify the perfect emotion. Our brain does this in a fraction of seconds. Of course, there can't be any such model as our brain but in the endeavour to make machines work like human brains, we need to person similar kind of activities. The speech signal has various types of features like spectral features, temporal features, and many more coming under the subcategories which we would be discussing in the latter part of this paper. These features are extracted from the speech signal are selected based on their impact and are loaded onto the model for training purposes.

Fig 1: Block Diagram of Speech Emotion Recognition Model

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 1300-1306 www.ijaem.net ISSN: 2395-5252

Now that we are clear with the idea of speech emotion recognition let's try to understand why is it needed.

Well, there are various reasons for the same. Due to the hectic lifestyle and stressful conditions,the depression rate is increasing. As per the reports of the World Health Organization [2]about 264 million people are facing some kind of mental disorders or are under depression. A major share of this comes under the age group of 15-29 yearsold. If the youth is depressed, we can't think of the growth and glory of the nation. Almost 8,00,000 people die due to suicide each year. Many of them indulge in committing suicide as they don't have anyone with whom they can speak out their heart or don't want others to know their condition. Many of them are not even aware that they are under depression. As these emotions can be identified by our model, we can further train it to respond and give emotional assistance to the person and thereby help him/her come out of depression.

Along with giving emotional assistance, this system can change the picture of reviewing system. We have sentiment analyzers for analyzing the text reviews over any product however a vocal review will always be the best way to communicate the review in the exact tone and mood of the speaker.

II. OVERVIEW OF SPEECH EMOTION RECOGNITION

2.1Database:

Taking about the workflow of the system, it begins with collecting the database. This database includes voice recordings of various actors that are collected together to form a huge database.

You can use datasets like:

1. RAVDESS: There are a total of 1500 audio recordings of 24 different actors which include 12 male and 13 female in different acted emotions.
2. SAVEE: This dataset includes around 500 audio recordings by 4 different male actors.

We can use any or both of these datasets to train our system however when it comes to testing identifying emotion gets difficult for natural emotions since actors act that particular emotion emphasizing all the features whereas a normal person may not always express the emotion with that impact.

Thus, the difficulty of identifying emotion increases from acted to natural database.

Overview of SER	Databases • Acted • Elicited • Natural
	Preprocessing • Framing • Windowing • Normalization • Noise Reduction • Feature Selection
	Features • Prosodic • Spectral • Voice Quality
	Classification • Classical Classifiers • Classifiers based on Deep learning • Deep Learning-Based Enhancement Techniques

Fig1. Overview of Speech Emotion Recognition

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 1300-1306 www.ijaem.net ISSN: 2395-5252

III. PREPROCESSING:

This is the very first step after the collection of data. This step includes various activities like dividing the signal into frames applying some functions to it, amplification, noise removal, etc.

2.1 Framing:

When talking about a long continuous speech signal, the emotions in the signal may vary in the frames of this signal. To avoid any such errors, framing also known as segmentation is carried out over the speech signals. In this process of framing, the continuous sound signal is cut into a smaller piece of a signal. These pieces are called frames. These frames are of fixed lengths. The frames of length as small as 20 to 30 ms generally have the same kind of emotion embedded in them which somehow smooths our task further.

2.2 Windowing:

As the next step in this process, we'll apply the window function to the frames that were created in the last step. This is done to avoid the leakages that may further occur due to Fast Fourier Transform (FFT). We mostly use a hamming window; whose window size is:

$$\omega(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{M-1}\right) \quad 0 \le n \le M-1$$

2.3 Voice Activity Detection:

If we consider any voice input, it always comprises three types of voices. We have voiced speech, unvoiced speech, and silence. This silence is heard because of no activity. It constitutes itself as a part of the noise in the signal. The second biggest noise constraint is this unvoiced signal. This is the part of speech that isn't deliberately created. This speech is added to the main voiced speech and causes noise. Removing this unvoiced speech and focusing only on the voiced speech is a big challenge in this process.

2.4 Noise Reduction:

Once we get all the input signals all we have to make them clear by removing the noise added to the sound signals. Also, the strength of the sound signal needs to be enhanced to make further operations comparatively easier. Few of the techniques here are used to normalize the variations between the sound and recording signal so that no further error occurs in the recognition process.

IV. FEATURE EXTRACTION

In this process, we shall deal with extracting the features from the pre-processed signal and storing them in a Comma Separated File(.csv). Before discussing how to extract the features it is very important to understand the features of a speech signal.

Speech being a continuous signal these features may vary with each frame. There are no specific sets of frames that are declared to be the correct set. We have a wide range of features that can be extracted using different algorithms and methodologies.

Let's discuss a few common types of features:

2.5 Prosodic Features:

This majorly stresses the usage of words in connection. These are the features that a human can detect easily. It majorly focuses on intonation and rhythm.

Example: Intonation, stress, rhythm

2.6 Spectral Features:

In this type, the features are determined considering the shape of the vocal tract while talking. There are various types of features here. Mel-frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC), Gammatone Frequency Cepstral Coefficients (GPCC), Lo-frequency Power Coefficients (LFCC) fall under this category.

2.7 Voice Quality Features:

These are the physical features of the vocal tract. They include the jitter, shimmer, harmonics to noise ratio. There is a strong correlation between these features and the emotions to be detected.

V. CLASSIFICATION:

We can try and test the accuracies using the various model. There are 3 categories.

5.1 Classical Classifiers:

This type of classifiers includes Support Vector Machine, Hidden Markov Model, Gaussian Mixture Model, Artificial Neural Networks, Decision Trees, Ensemble Methods.

5.2 Classifiers based on Deep learning:

This category includes Convolutional Neural Network, Deep Neural Network, Recurrent Neural Networks, Long-short Term Memory Networks.

We have worked over this category in this paper and further, we shall compare their results and efficiencies as well.

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 1300-1306 www.ijaem.net ISSN: 2395-5252

5.3 Deep Learning-Based Enhancement Techniques:

This category includes autoencoders, multitask learning, attention mechanism, transfer learning, adversarial training.

Further, we will be looking forward to the actual implementation, the accuracies, and the shortcomings of each model concerning SER.

VI. METHODS:

We have tried and tested various models to check the results. For the same purpose, we will be dealing with the deep learning models. As RNN and CNN are used for speech processing, we started our procedure with RNN first.

Before that, we need to extract the features from the audio file. For the same purpose, we have used librosa library in python. Using librosa we can extract all the spectral features.

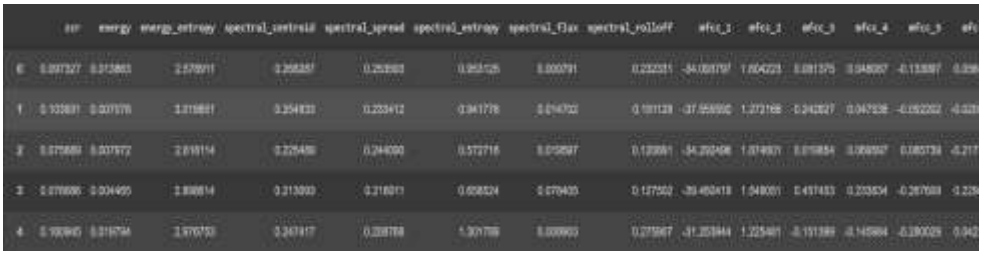

Fig2. Extracted features

There are a total of 68 features out of which 40 important features are selected and are sent to the model for further testing.

Features are as follows:(refer Fig3)

Fig3. Total Features

6.1 MLPClassifier:

Working with our first classifier the MLP Classifier.

MLPClassifier i.e., Multi-layer Perception classifier which depends on an underlying Neural Network to perform classification.

Using MLP we tried building the classification model however its accuracy was not as per the expectations.

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 1300-1306 www.ijaem.net ISSN: 2395-5252

```
/usr/local/lib/python3.7/dist-packages/sklearn/neural_network/_multilayer_perceptron.py:571: Co
  % self.max_iter, ConvergenceWarning)
GridSearchCV(cv=3, error_score=nan,
             estimator=MLPClassifier(activation='relu', alpha=0.0001,
                                     batch_size='auto', beta_1=0.9,
                                     beta_2=0.999, early_stopping=False,
                                     epsilon=1e-08, hidden_layer_sizes=(100,),
                                     learning_rate='constant',
                                     learning_rate_init=0.001, max_fun=15000,
                                     max_iter=100, momentum=0.9,
                                     n_iter_no_change=10,
                                     nesterovs_momentum=True, power_t=0.5,
                                     random_state...
                                     solver='adam', tol=0.0001,
                                     validation_fraction=0.1, verbose=False,
                                     warm_start=False),
             iid='deprecated', n_jobs=-1,
             param_grid={'activation': ['tanh', 'relu'],
                         'alpha': [0.0001, 0.05],
                         'hidden_layer_sizes': [(50, 50, 50), (50, 100, 50),
                                                (100,)],
                         'learning_rate': ['constant', 'adaptive'],
                         'solver': ['sgd', 'adam']},
             pre_dispatch='2*n_jobs', refit=True, return_train_score=False,
             scoring=None, verbose=0)
```

Fig4. Overview of MLP Classifier.

The accuracy of this model is 42% which is not even close to our expected results.

6.2 Recurrent Neural Network:

Further moving to the Recurrent Neural Network model, as RNN is a good choice for speech processing we tried to train our model with RNN.

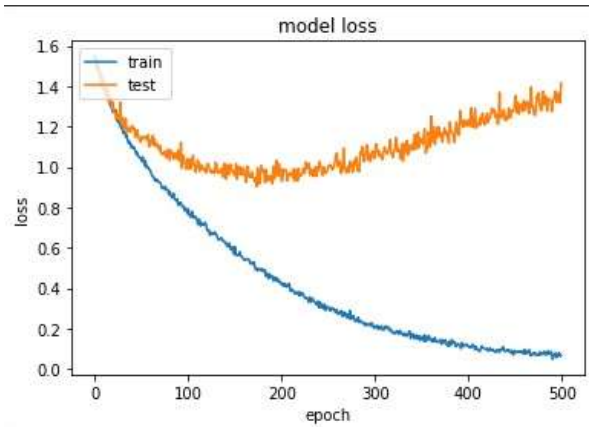

Fig5. Overview of RNN.

Though the model gives an accuracy of 99% its of no use to us since the loss is high as it is an overfitting model so we need to try some other classification technique in order to get the best possible results.

6.3 Convolutional Neural Network:

Our next attempt was with Convolutional Neural Network. CNN is feed forward neural network with some variations of multilayer perceptron that are designed to use minimal amounts of preprocessing.

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 1300-1306 www.ijaem.net ISSN: 2395-5252

```
Model: "sequential_4"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_8 (Conv1D)            (None, 40, 128)           768
_________________________________________________________________
activation_12 (Activation)   (None, 40, 128)           0
_________________________________________________________________
dropout_8 (Dropout)          (None, 40, 128)           0
_________________________________________________________________
max_pooling1d_4 (MaxPooling1 (None, 5, 128)            0
_________________________________________________________________
conv1d_9 (Conv1D)            (None, 5, 256)            164096
_________________________________________________________________
activation_13 (Activation)   (None, 5, 256)            0
_________________________________________________________________
dropout_9 (Dropout)          (None, 5, 256)            0
_________________________________________________________________
flatten_4 (Flatten)          (None, 1280)              0
_________________________________________________________________
dense_4 (Dense)              (None, 7)                 8967
_________________________________________________________________
activation_14 (Activation)   (None, 7)                 0
=================================================================
Total params: 173,831
Trainable params: 173,831
Non-trainable params: 0
_________________________________________________________________
```

Fig6. Summary of CNN.

On checking upon the accuracy of this model we reach a point where the model is quite stable and accurate.

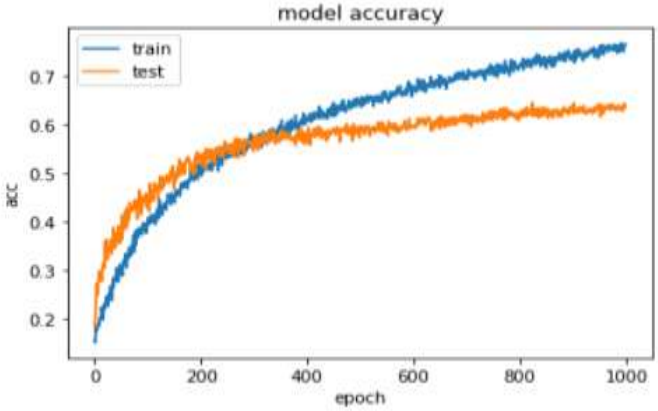

Fig7. Result of CNN.

The graph shows us the test and train accuracy of the model.
For more details of the model visit https://github.com/sarang109/Speech-Emotion-Recognition.

VII. RESULTS:

The accuracy of the Convolutional Neural network model is found to be 63.71% and the results are quite good as well.

VIII. CONCLUSION:

This paper has explained a lot about speech emotion recognition using deep learning techniques. Deep learning topics like Convolutional Neural Network, Recurrent Neural Network, are under focus for many years due to their layered architecture. This layer-wise architecture proves to be fruitful for identify the emotions of the speaker. Though there are a few limitations the system performs well. Our system recognizes the basic 7 emotions namely, anger, happy, neutral, disgust, sad, fear, surprise however when a person talks with mixed emotions it gets difficult to understand the right emotion. Also, when tested with a natural dataset, identification becomes difficult. Overall, Speech Emotion Recognition can be implemented successfully using deep learning.

REFERENCES

[1]. R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "**Speech Emotion Recognition Using Deep Learning Techniques: A Review**," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.

[2]. Mehmet Berkehan Akçay, Kaya Oğuz, **Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Communication**, Volume 116, 2020, Pages 56-76, ISSN 0167-6393, https://doi.org/10.1016/j.specom.2019.12.001.

[3]. I. Shahin, A. B. Nassif and S. Hamsa, "**Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network**," in IEEE Access, vol. 7, pp. 26777-26787, 2019, doi: 10.1109/ACCESS.2019.2901352.