ISSN No:-2456-2165

Speech Emotion Recognition using Deep Learning

Akash Raghav, Dr. C. Lakshmi
Computer Science & Engineering SRM IST
Chennai, India

Abstract:- The goal of the project is to detect the speaker's emotions while he or she speaks. Speech generated under a condition of fear, rage or delight, for example, becomes very loud and fast, with a larger and more varied pitch range, However, in a moment of grief or tiredness, speech is slow and low-pitched. Voice and speech patterns can be used to detect human emotions, which can help improve human-machine interactions. We give Deep Neural Networks CNN, Support Vector Machine, and MLP Classification based on auditory data for emotion produced by speech, such as Mel Frequency Cepstral Coefficient classification model (MFCC).Eight different emotions have been taught to the model (neutral, calm, happy, sad, angry, fearful, disgust, surprise), Using the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset as well as the TESS (Toronto Emotional Speech Set) dataset, we found that the proposed approach achieves accuracies of 86 percent, 84 percent, and 82 percent, respectively, for eight emotions using CNN, MLP Classifier, and SVM Classifiers.

I. INTRODUCTION

The most common method of human interactions is via spoken language is the cornerstone for the purpose of information sharing& has been an important aspect of society from the dawn of time. Emotions, on the other hand, can be traced back to basic instinct before the emergence of the modern spoken language, and can be regarded one of the first forms of natural communication. It is also used in a variety of practical applications in many sectors such as BPO-Business Process Outsourcing Centers& call centers to analyze emotion essential for determining client pleasure.

Proposed Model depicts how each of us displays more than one basic emotion at a time, but we believe that recognizing which emotion & analyzing what percentage of the emotions are jumbled is exceedingly strenuous for both the speaker and the listener. According to this scenario, it is decided to create a model that only recognizes the emotions in the audio track that are more valuable.Various methods, such as computer vision or text analytics, have been tried to classify feelings by a machine. The purpose of this study is to employ Mel-frequency cepstral Coefficients(MFCC) with pure audio data.

Speech is one of the most basic human capacities, allowing us to connect with one another, express ourselves, and, most importantly, give us a feeling of self. It is one of the most important aspects of mental and physical health. Because emotions can influence how we act and interpret events, a speaker's speech transmits both its linguistic meaning and the emotion with which it is given.

II. LITERATURE REVIEW

Many classification algorithms have been presented in this field of research in recent years. Iqbal et al.[6], however for the sake of this paper, we only looked at the work done on RAVDESS. Iqbal etal.[6]developed a granular classification technique that merges Gradient Boosting, KNN, and Support Vector Machineto work on the RAVDESS dataset used in this study with roughly 40% to 80% overall accuracy ,depending upon the tasks. In different datasets, the proposed classifiers performed differently. He created three datasets: one with only male recordings, one with exclusively female recordings, and one with both male and female recordings. In RAVDESS, Support Vector Machine and KNN show 100 percent accuracy in both angry and neutral situations (male),however gradient boosting outperforms support vector machine and KNN algorithm in happiness and sadness. Support Vector machine in RAVDESS (female) achieves a level of accuracy of 100 percent in fury, similar to the male counterpart. Except for sadness, Support Vector machine has a decent overall performance in general. KNN also performs well in rage and neutral situations, scoring 87 percent and 100 percent, respectively.

Gradient Boosting performs poorly in anger and IEEE (978-1-7281-4384-2/20/$31.00 © 2020) neutral. In comparison to other classifiers, KNN performs poorly in happiness and sadness. SVM and KNN perform far better in rage and neutral than Gradient Boosting in a mixed male and female sample. In both happiness and despair, KNN's performance is dreadful. With the exception of SVM, The male dataset has superior average classifier performance than the female dataset. SVM has a higher accuracy in a combined database Than gender based dataset.

Another technique developed by Jannat et al.[7] achieved 66.41 percent accuracy on audio data and above 90% accuracy when integrating audio and visual data.. Faces and audio waveforms, in particular, are present in the preprocessed image data.

Xinzhou Xu [3] et al. modified the Spectral Regression model by utilizing the joins of ELMs – Extreme Learning Machines and SL-Subspace Learning in order to overcome the drawbacks of spectral regression-based GE-Graph Embedding and ELM. We have to correctly describe these relationships among data using the GSR model in the execution of the SER- Speech Emotion Recognition .For the same, many embedded graphs were created. The impact and practicality of the strategies were determined by Demonstrationover4Speech Emotional Corpora when compared to past methods such as ELM and Subspace Learning (SL) methodologies. Exploring embedded graphs at a deeper level can help the system produce better results.

ISSN No:-2456-2165

Only Least-Square Regression and l2-norm minimization were used in the regression stage.

To detect speech depression, Zhaocheng Huang[4] et al deploy a heterogeneous token-based method. Acoustic areas and abrupt shifts are only and collectively determined in junctions between distinct embedding methods. Contributions to the detection of depression, as well as numerous health issues that may impair vocal generation, were utilised. Landmarks are used to retrieve data particular to each type of articulation at a given point in time. This system is a mix of the two. LWs and AWs hold a wide range of information. LWs depict the sudden variations in speech articulation on the contemporary, while AW holds a part of acoustic space in a single token each frame. The hybrid join of the LWs and AWs allows for the exploration of numerous aspects, including articulatory dysfunction as well as traditional acoustic features.

For cross corpus voice recognition, Peng Song [5] proposes the Transfer Linear Subspace Learning (TLSL) paradigm. TULSL, TSLSL, and TLSL methods were all counted. The goal of TLSL is to extract robust character representations from corpora into a trained estimated subspace. TLSL improves on the currently utilized transfer learning algorithms, which merely seek for the most portable features components. TLSL achieves even better results than the six baseline approaches with statistical significance, and TSLSL achieves even better results than TULSL; in fact, all transfer learning techniques are more accurate than traditional learning procedures. The good transfer learning approaches based on characteristics transformation, such as TLDA, TPCA, TNMF, and TCA, are greatly outperformed by TLSL. One of the major drawbacks of these early transfer learning methods is that they focus on finding the portable components of traits while ignoring the less informative sections. When it comes to transfer learning results, the less informative bits are also important. TLSL is used for cross-corpus identification of speech emotion.

Jun Deng [6] et al. focused on unsupervised learning with automatic speech emotion encoders in this research. To combine generative and discriminative training, partially supervised learning approaches designed for situations with non-labeled data were applied. Five databases in various situations were used to test the procedure successively. In settings with a reduced number of labeled instances, the suggested approach improves recognition performance by acquiring prior knowledge from non-labeled data. These techniques may deal with a wide range of problems and incorporate knowledge from different fields into classifiers, resulting in excellent performance. This shows that the model can distinguish speech emotions from a mix of labeled and unlabeled data.The residual neural network revealed that dense architectures enable the classifier to extract complex structures in image processing.

Ying Qin[7] et al. presented a Cantonese-speaking PWA narrative speech that serves as the foundation for a fully automated assessment system. Experiments based on the recommended data on text characteristics may be able to detect linguistic impairment in aphasic speech. The Siamese network learned text qualities that were highly linked with the AQ scores. The confusion network was built using the improvised representation of ASR output, and the robustness of text characteristics was praised. There was a pressing need to improve ASR's performance on aphasic speech in order to generate speech with more robust qualities. To use this proposed methodology, databases of disordered speech and other languages were required. As shown in clinical practice, the most ideal one is automatic classification of aphasia variants, which necessitates a significant amount of data collection.

III. IMPLEMENTATION DETAILS

A. Methodology

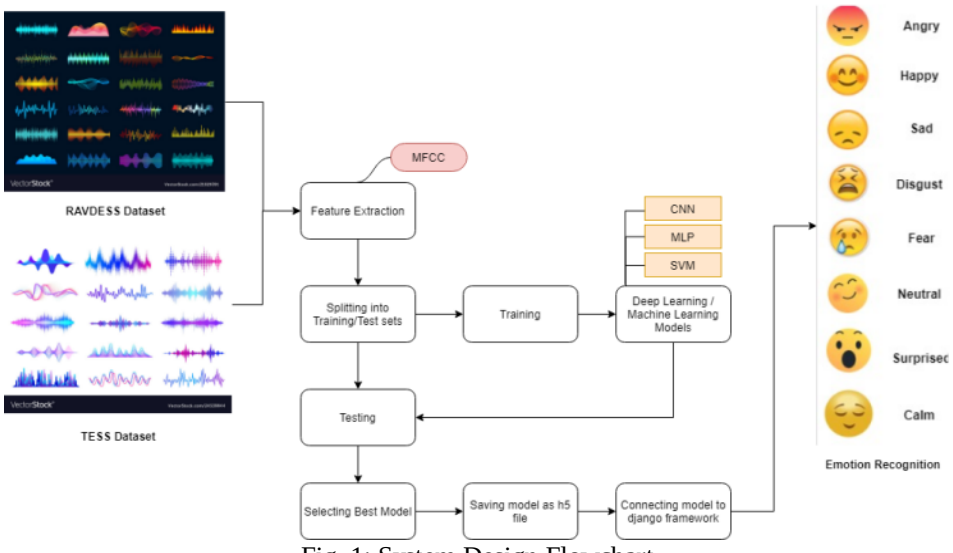

Fig. 1: System Design Flowchart

ISSN No:-2456-2165

The emotion recognition classification models shown here use a deep learning strategy that employs CNN, SVM, and MLP classifiers. MFCC is the single feature required to train the model, which is also known as the "spectrum of a spectrum."

The best sound formalization method for automatic speech recognition tasks has been proven to be MFCC, which is an MFC-Mel frequency variant Because the MFC coefficients have been frequently employed because of their ability to convey the amplitude spectrum of a sound wave in a compact vectorial form.

To obtain statistically steady waves, The audio files are divided into frames, which are normally determined by the size of a set window. The frequency scale of the "Mel" is reduced to standardize the amplitude spectra. This procedure is carried out in order to identify with the frequency in order to precisely replicate the wave using the human auditory system.

A total of 40 features have been retrieved from each audio file. To build the feature,each audio file was transformed into a floatingpoint time series. The time series was then converted into an MFCC sequence.The MFCC array on the horizontal axis is transposed , and arithmetic mean is calculated .

B. Dataset

This task's dataset consists of 5252 samples collected from the following sources:

- RAVDESS – Ryerson Audio Visual Database of Emotional Speech and Song
- TESS – Toronto Emotional Speech Set

Sample consists of:

RAVDESS - 1440 speech files and 1012 song files.The dataset consists of recordings of 24 (12-females & 12-males) professional actors who speak in neutral north American accent. Happy, angry, calm, sad, afraid, disgust, and surprise expressions can be found in speech, whereas happy, angry, calm, fearful, and sad emotions can be found in song.

TESS - 2800 files. This dataset consists of two actresses aged between 26 and 64 recited around 200 set of taeget phrases in the carrier phrase"Say the word _____," and Each of the seven emotions were recorded on the set (anger, neutral, fear, pleasant surprise, disgust, sadness, and happiness). There are 2800 stimuli in all. Two actresses from the Toronto area have been chosen. Both actors are native English speakers with a university education and musical background. Both actresses' audiometric thresholds are within the normal range, according to audiometric testing.

0 - Neutral
1 - Calm
2 - Happy
3 - Sad
4 - Angry
5 - Fearful
6 - Disgust
7- Surprised

These are classes the model aims to predict. There is no calm class in TESS, hence this dataset is distorted. As a result, there is less data for that particular class, as evidenced by the classification report.

- **Identifiers in filename -**

Modality – 01 = Full-AV
02 = Video-only
03 = Audio-only

Vocal Channel – 01 = Speech
02 = Song

Emotion – 01 = Neutral 05 = Angry
02 = Calm 06 = Fearful
03 = Happy 07 = Disgust
04 = Sad 08 = Surprised

Emotional Intensity – 01 = Normal
02 = Strong

Statement – 01 = “Kids are talking by the door”
02 = “Dogs are sitting by the door”

Repetition – 01 = 1st repetition
02 = 2nd repetition

Actor – 01 to 24
Odd-numbered are male
Even-numbered are female

C. Algorithms

- The Classification task's deep neural network (CNN) is operationally reported in Fig. 1. For each and every audio file that has been given as an input, the network can function with 40-feature vectors. The 40 values reflect the compressed numerical representation of a two-second audio frame. As a result, A set of 40x1 training files were used to run one round of a 1D CNN using a ReLU activation function, a 20% dropout, and a 2 x 2 max-pooling function.

The ReLU(rectified linear unit) is formalised as g(z) = max0, z and allows us to get a large value in the event of activation by representing hidden units with this function. In this situation, Pooling can let the model focus primarily on the most relevant characteristics of each piece of input, resulting in position invariant results. We repeated the procedure, this time adjusting the kernel size. After that, Another dropout was applied, and the result was flattened to make it compatible with the subsequent layers.
Finally, for each of the properly encoded classes, we calculated the probability distribution.

Neutral - 0
Calm - 1
Happy - 2
Sad - 3
Angry - 4
Fearful - 5
Disgust - 6
Surprised – 7

ISSN No:-2456-2165

Using a softmax activation function on one Dense layer (fully connected layer).

```
In [ ]: model.summary()

Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_1 (Conv1D)            (None, 40, 64)            384
_________________________________________________________________
activation_1 (Activation)    (None, 40, 64)            0
_________________________________________________________________
dropout_1 (Dropout)          (None, 40, 64)            0
_________________________________________________________________
max_pooling1d_1 (MaxPooling1 (None, 10, 64)            0
_________________________________________________________________
conv1d_2 (Conv1D)            (None, 10, 128)           41088
_________________________________________________________________
activation_2 (Activation)    (None, 10, 128)           0
_________________________________________________________________
dropout_2 (Dropout)          (None, 10, 128)           0
_________________________________________________________________
max_pooling1d_2 (MaxPooling1 (None, 2, 128)            0
_________________________________________________________________
conv1d_3 (Conv1D)            (None, 2, 256)            164096
_________________________________________________________________
activation_3 (Activation)    (None, 2, 256)            0
_________________________________________________________________
dropout_3 (Dropout)          (None, 2, 256)            0
_________________________________________________________________
flatten_1 (Flatten)          (None, 512)               0
_________________________________________________________________
dense_1 (Dense)              (None, 8)                 4104
_________________________________________________________________
activation_4 (Activation)    (None, 8)                 0
=================================================================
Total params: 209,672
Trainable params: 209,672
Non-trainable params: 0
_________________________________________________________________
```

```
In [ ]: loss, acc = new_model.evaluate(x_testcnn, y_test)
        print("Restored model, accuracy: {:5.2f}%".format(100*acc))

        1734/1734 [==============================] - 0s 95us/step
        Restored model, accuracy: 85.47%
```

Table 1: Model Summary

- The Multilayer Perceptron (MLP) is a forward-feeding artificial neural network (ANN). Back propagation is used by MLP during training as a supervised learning strategy. MLP is distinguished from a linear perceptron by its numerous layers and non-linear activation. It has the ability to identify data that isn't linearly separable.

```
In [ ]: accuracy=accuracy_score(y_true=y_test, y_pred=y_pred)
        print("Accuracy: {:.2f}%".format(accuracy*100))

        Accuracy: 83.32%
```

```
In [ ]: from sklearn.metrics import classification_report
        print(classification_report(y_test,y_pred))
```

	precision	recall	f1-score	support
angry	0.94	0.84	0.89	199
calm	0.72	0.72	0.72	86
disgust	0.87	0.82	0.84	142
fearful	0.86	0.83	0.85	176
happy	0.77	0.84	0.80	186
neutral	0.72	0.96	0.82	165
sad	0.89	0.80	0.85	199
surprised	0.89	0.79	0.84	160
accuracy			0.83	1313
macro avg	0.83	0.83	0.83	1313
weighted avg	0.84	0.83	0.83	1313

Table 2: MLP model result on the test

- Support Vector Machine (SVM) is a supervised machine learning technique for solving classification and regression issues. It is, however, mostly used to tackle categorisation questions. Every data item is represented as a point in n-dimensional space (where n denotes the number of features), with the value of each feature being the SVM algorithm's value for a given coordinate. To avoid attributes in higher numeric ranges while processing data, Before using it with an SVM classifier, it might be scaled.Scaling also helps to prevent some arithmetic issues during the calculating process.

	precision	recall	f1-score	support
angry	0.89	0.92	0.91	153
calm	0.62	0.94	0.75	77
disgust	0.81	0.91	0.86	119
fear	0.76	0.80	0.78	142
happy	0.94	0.73	0.82	164
neutral	1.00	0.78	0.88	108
sad	0.77	0.78	0.77	156
surprised	0.83	0.80	0.81	132
accuracy			0.82	1051
macro avg	0.83	0.83	0.82	1051
weighted avg	0.84	0.82	0.82	1051

----accuracy score 82.30256898192198 ----

Table 3: SVM model result on the test

IV. RESULT AND DISCUSSION

On the RAVDESS and TESS datasets, when compared against baselines and the state of the art, the evaluation results show that the model is effective.

For each of the emotional classifications, Table I illustrates the precision, recall, and F1 values obtained. The results show that precision and recall are very well balanced, allowing us to acquire F1 values that are evenly dispersed around the value 0.85 for practically all classes. The model's robustness is demonstrated by the limited range of F1 values, which efficiently classify emotions into eight separate categories. The model is less accurate in the classes "Calm" and "Disgust," which is understandable given that they are the most difficult to discern not only by speaking but also by observing facial expressions or evaluating written language, as indicated in the Introduction.

```
In [ ]: from sklearn.metrics import classification_report
        report = classification_report(new_Ytest, predictions)
        print(report)
```

	precision	recall	f1-score	support
0	0.88	0.91	0.89	190
1	0.77	0.76	0.76	117
2	0.90	0.84	0.87	266
3	0.80	0.86	0.83	246
4	0.89	0.88	0.88	265
5	0.88	0.80	0.84	246
6	0.81	0.92	0.86	202
7	0.88	0.85	0.86	202
accuracy			0.85	1734
macro avg	0.85	0.85	0.85	1734
weighted avg	0.86	0.85	0.85	1734

Table 4: CNN model result on the test

ISSN No:-2456-2165

We decided to examine the findings acquired from two additional methods, namely SVM and MLP classifier, to determine the efficacy of the emotion classification provided in this study.

On all classes, our model's F1 values outperform baselines and competition, as shown in Tab II. However, it is important to note that the performance loss is minor and was implemented to avoid overfitting. It's common knowledge that as the number of classes increases, the categorization task becomes more complex and inaccurate.

CLASS	MLP	SVM	CNN
SAD	0.81	0.82	0.80
ANGRY	0.89	0.91	0.89
HAPPY	0.82	0.84	0.90
DISGUST	0.80	0.81	0.81
SURPRISE	0.80	0.87	0.88
NEUTRAL	0.89	0.93	0.88
CALM	0.75	0.64	0.77
FEAR	0.84	0.81	0.88

Table 5: Each class's F1-score is compared to the baselines (SVM, MLP)

Nonetheless, CNN-MFCC model provided here achieves an F1 score that is comparable to the two jobs we were given. Figures 2 and 3 show another indicator of model reliability. Up to the 200th epoch, the value of loss (model accuracy error) on both the test and training sets tends to decrease. From the 100th epoch forward, the decline is less noticeable, but it is still noticeable.

In Fig. 3, the average value of accuracy across all classes is shown, which, in contrast to the loss, increases as the age of the child increases. These figures are nearly identical between the training and test datasets, demonstrating that the model was not overfitted during training. The results are consistent with the previously discovered F1 scores.

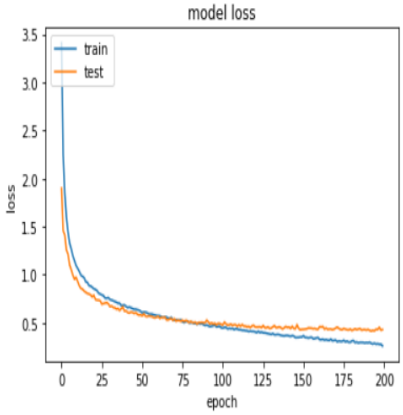

Fig. 2: Cost Function

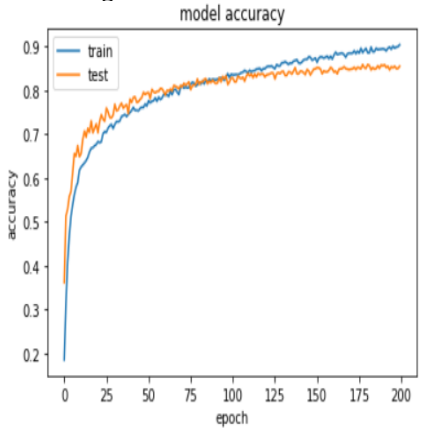

Fig. 3: Accuracy

V. CONCLUSION AND FURTHER ENHANCMENTS

In the above paper , We used audio recordings from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Toronto Emotional Speech Set (TESS) to show an architecture based on deep neural networks for emotion categorization .The model is trained to classify 7 different emotions (neutral, calm, happy, sad, furious, terrified, disgusted, startled) and obtained an overall F1 score of 0.85, with the best results in the Happy class (0.90) and the worst results in the silent class (0.80). (0.85). (0.85). (0.77).To obtain this outcome, we extracted the MFCC features (spectrum of a spectrum) from the audio recordings used for training. Using 1D CNNs, max-pooling operations, and Dense Layers, we trained a deep neural network to consistently predict the probability of distribution of annotation classes based on the aforementioned representations of input data.The method was put to the test using data from the RAVDESS dataset. With an average F1 score of 0.84, we employed an MLP Classifier trained on the same dataset.across the eight classes as a baseline for our task. Following the MLP classifier, we trained an SVM classifier that scored 0.82 on the F1 scale. On the test set, Our final option was a deep learning model with an F1 score of 0.86. The positive results show that deep neural network-based techniques provide an outstanding foundation forsolving problem. They are generic enough in particular to perform correctly in a real-world application setting, Earlier versions of this paper solely used the RAVDESS dataset, with TESS being added afterwards. Additionally, previous versions of this research

used audio features from the films in the RAVDESS dataset. Because it was shuffling very similar files in the training and test sets, this component of the pipeline was removed, enhancing the model's accuracy (overfitting).

REFERENCES

[1.] Y. Chen, Z. Lin, X. Zhao, S. Member, G. Wang, and Y. Gu, "Deep Learning-Based Classi fi cation of Hyperspectral Data," pp. 1–14, 2014.
[2.] Jannat, R., Tynes, I., Lime, L. L., Adorno, J., and Canavan, S. Ubiquitous emotion recognition using audio and video data. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (2018), ACM, pp. 956–959.
[3.] X. Xu, J. Deng, E. Coutinho, C. Wu, and L. Zhao, "Connecting Subspace Learning and Extreme Learning Machine in Speech Emotion Recognition," IEEE, vol. XX, no. XX, pp. 1–13, 2018.
[4.] Logan, B., et al.: Mel frequency cepstral coefficients for music modeling. In ISMIR (2000), vol. 270, pp. 1–11.
[5.] Z. Huang, J. Epps, D. Joachim, and V. Sethu, "Natural Language Processing Methods for Acoustic and Landmark Event-based Features in Speech-based Depression Detection," IEEE J. Sel. Top. Signal Process., vol. PP, no. c, p. 1, 2019.
[6.] Nair, V., and Hinton, G. E. : Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning (ICML-10) (2010), pp. 807–814
[7.] J. Deng, X. Xu, Z. Zhang, and S. Member, "Semi-Supervised Autoencoders for Speech Emotion Recognition," vol. XX, no. XX, pp. 1–13, 2017.