Speech Emotion Recognition Using 1D-CNNs with MFCCs

1. Abstract:

Speech Emotion Recognition (SER) has gained a lot of attention in recent years. Due to the advancement in the field of audio signal processing and increased widespread usage of voice-based assistants, we can find applications of Audio-AI everywhere. But, the problems that still remain with all the voice assistants and other voice-based solutions is that they aren't able to understand the emotions of the speaker. This increases the complexity required to understand the intent of the user. Therefore, in this research work, we have proposed a deep learning-based method for Speech Emotion Recognition. Our proposed model is trained on a combination of three datasets - RAVDESS speech, RAVDESS song and TESS dataset. It utilizes 1D CNNs for identifying the emotion of the speaker through his voice. Our model has shown great performance in the task of Speech Emotion Recognition with an accuracy of 86.55% and weighted F1-Score of 86.62%.

2. Introduction:

Speech Emotion Recognition (SER) is the technique of recognizing the emotional aspects of speech irrespective of semantic material. While humans can effectively recognize the emotions as a natural part of speech communication, the ability to incorporate it automatically using programmable devices is still an ongoing subject of research. Automatic Emotion Recognition Systems studies aim to establish efficient, real-time applications for detecting the emotions of mobile phone users, call centre operators and consumers, car drivers, pilots and many other users of human-machine communication. The use of robot emotions has been accepted as a central factor in making robots look and work in a human-like manner.

Robots who are capable of feeling emotions can have an appropriate emotional response and exhibit emotional personality. In certain situations, humans can be replaced by computer-generated characters with the ability to engage in very believable and persuasive

encounters by appealing to human emotions. Machines ought to consider the emotions conveyed by speech. Only with this power can a genuinely substantive dialogue be created, focused on mutual human-machine trust and understanding.

Speech Emotional Recognition is an important topic that is of growing concern to researchers due to its various uses, such as audio monitoring, e-learning, health trials, lies identification, entertainment, video games and call centres. However, this issue remains a big obstacle for sophisticated machine learning techniques. One of the reasons for this modest success is the difficulty of selecting the best features. Besides, the presence of background noise in audio samples, such as real-world voices, may have a dramatic influence on the success of the machine learning model[1]. However, the development of good emotional speech recognition models could greatly boost user experience in applications involving human-machine interactions, for example in the fields of Artificial Intelligence (AI) or Mobile Health (mHealth). Indeed the ability to identify emotions from audio recordings and thus the ability to mimic these emotions may have a major effect on the AI field. Various virtual assistants in the field of mHealth may dramatically increase their efficiency by using certain models. Furthermore, emotional speech recognition devices are unpretentious in terms of hardware specifications.

Deep learning models are currently being used to solve recognition problems such as facial recognition, voice recognition, image recognition, and speech recognition[3–6]. One of the primary advantages of deep learning techniques is the automatic selection of attributes that may be applied, for example, to the important qualities inherent in sound files that have unique emotions in the task of understanding speech emotions.

Speech emotion recognition can be defined as the detection of emotional states (features) of a speaker from speech and extraction of selected features for the final classification. In some cases, the process is assisted by the speech recognition system that contributes to the classification using linguistic information. These two techniques deal with a challenging task since emotional states are different from person to person. Speech emotion recognition systems can be helpful in intelligent assistance [7], criminal investigation [8], health care systems [9], and many more domains.

Emotion recognition in spoken dialogues has been gaining increasing popularity over the years. Speech Emotion Recognition (SER) is a hot research study. Topic in the field of Human-Computer Interaction (HCI). It has theoretically wide-ranging applications, such as robot interfaces, accounting, call centres, cardboard networks, video games, etc. In the context of classroom orchestration or e-learning, information on the emotional state of students may include a perspective on learning and Improving the standard of instruction. For example, teachers can use SER to decide which subjects can be taught and should be able to establish techniques for controlling emotions in the learning environment. That's why the mental status of the learner should be recognized in the classroom. In general, the SER is a computational task consisting of two key components: the extraction function.

Recognition of feelings has recently become an important focus of study. There are already a few models that can predict emotion, but the purpose of this paper is to construct a specific model that is not only lightweight but also fast and precise. The identification of emotions from human expression is important to the comprehension of human behaviour. For any computer to correctly decipher the desired meaning in the voice, it must grasp the emotion of the spoken word. Emotions influence speech modulations, and these modulations can also alter the background. The goal of this paper is to suggest a device that can accurately detect emotions from expression. The area of human speech recognition of emotions is very complex due to extremely overlapping regions of emotions, and it often becomes very difficult to differentiate between two emotions based solely on voice. Such ambiguity in the label assignment is responsible for low classification accuracy in existing systems. In the proposed system, we have worked on finding both the suitable feature set as well as the classifier.

3. Related Works:

Study in the field of human speech emotions is on the rise day by day. Too many new technologies are occurring every day because the science world has been able to leap forward and close the gap between machines and human understanding. Researchers have proposed various theories, methodologies and models for the successful detection of emotions in the acoustic field.

Neiberg et al.[10] concentrated on the recognition of random emotions. According to them, it is more difficult to classify emotions from a live or natural voice than to use a pre-recorded dataset. To explain sentiment in spontaneous real-time speech, they introduced a process by which they used three classifiers and averaged their results. They used the MFCC, the Low MFCC in the 20–300 Hz range and the pitch to perform the task. The findings indicate that two MFCCs have the same effects, whereas the MFCC is low above the pitch. The Gaussian mixture model (GMM) was used as a frame stage classifier.

Chenchah and Lachiri[11] have studied the effectiveness of Mel-Frequency Cepstral Coefficients (MFCC) and Linear Frequency Cepstral Coefficient (LPCC) in the detection of emotions using Hidden Markov Model (HMM) and Vector Support Devices (SVM). The model developed, produced 61 per cent accuracy with the Surrey Audio-Visual Expressed Emotion (SAVEE) database. Parthasarathy and Tashev[12] compared DNNs, RNNs and 1D-CNN models with MFCC features from a Chinese language dataset. They also attained 56 per cent accuracy for the CNN 1D model.

CNNs have achieved significant outcomes in identification practices ranging from image, video and sound. Wunarso and Soelistio[13] have developed a speech database (I-SpeED) based on the native Indonesian language. They extracted amplitude, speech duration, and approximation coefficients as inputs dependent on SVM. After their analysis and estimate, they obtained an average accuracy of 76.84%. Zamil et al.[14] proposed that the Mel Frequency Cepstrum Coefficient (MFCC) function be derived from speech signals to distinguish particular speech emotions. The derived attributes have been used to classify different emotions using the LMT classifier.

Using a voting system on the categorized frames, the speech signal emotion was observed. Two speech corpus, Emo-DB and RAVDESS, were used to test the model. Among the qualified models, the highest response accuracy was 70 per cent for seven different emotions. Yoon et al.[15] applied a multimodal approach to speech emotion detection. They used audio and text data from the IEMOCAP[16] database for this mission. The authors used MFCCs and text

tokens as input features for their model. The multimodal method resulted in a 71.8% accuracy of the test sample.

4. Material and Methods:

a. Dataset

We have used two datasets in this work:

- **Ryerson Audio-Visual Database for Emotional Speech and Song -** RAVDESS dataset[17] includes 7356 files made by 24 trained actors (12 female, 12 male), vocalizing two lexically matched comments in a neutral North American accent. Speech incorporates calm, happy, sad, angry, fearful, surprise and disgusted emotions, and the songs comprise calm, happy, sad, angry, and fearful emotions. Each expression appears at two degrees of emotional strength (normal, strong) with an additional neutral expression. All parameters are available in three formats: audio-only (16bit, 48kHz.wav), audio-video (720p H.264, AAC 48kHz,.mp4), and video-only (no sound). Ratings were given by 247 individuals who were representative of untrained adult study participants from North America.

  Each of the 7356 RAVDESS files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 02- 01-06-01-02-01-12.mp4). These identifiers define the stimulus characteristics:

  - Modality (01 = full-AV, 02 = video-only, 03 = audio only);
  - Vocal channel (01 = speech, 02 = song);
  - Emotion (01 = neutral, 02 = calm, 03 = happy, 04= sad, 05 = angry, 06 = fearful, 07 = disgust, 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 actors are male, even-numbered actors are female).

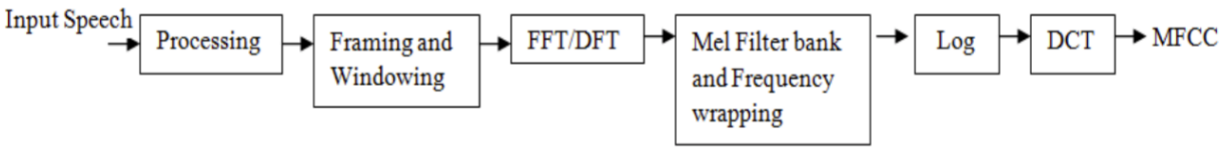
  NOTE: There is no strong intensity for the ’neutral’ emotion.

Speech signals have a different amplitude as each signal incorporates different emotions. It is obvious that a signal with a higher amplitude is likely to include anger. In other languages, this might be different. Emotions, happy and sad, have a common amplitude, and it is fascinating to differentiate between the two emotions using machine learning and to know which features are relevant.

- **Toronto Emotional Speech Set (TESS) Collection -** There are a set of 200 target words were spoken in the carrier phrase "Say the word _' by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 data points (audio files) in total[18]. The dataset is organised such that each of the two female actors and their emotions is contained within its folder. And within that, all 200 target words audio files can be found. The format of the audio file is a WAV format.

b. MFCC

Mel-frequency Cepstrum coefficient is the most used representation of spectral property of voice signals. These are the best for speech recognition as it takes human perception sensitivity with respect to frequencies into consideration. For each frame, the Fourier transform and the energy spectrum were estimated and mapped into the Mel-frequency scale. Mel-Frequency Cepstral Coefficients (MFCC) represents the spectral property of the speech signals. The vocal cords, the tongue, and the teeth are all the elements that filter the sound and make it unique for each speaker. The voice is determined by the shape of these elements. The shape manifests in the envelope of the short-time power spectrum, and MFCCs represent this envelope [19]. Usually, the process of calculating MFCC is shown in the figure below.



5. Method Proposed:

In this section, we introduce our proposed method based on 1D CNN for Speech Emotion Recognition and the method of feature extraction.

A. Data Pre-processing

For making a generalized model we combined RAVDESS speech, RAVDESS song and TESS dataset. This helped us in creating a model which can perform well in various scenarios. The number of instances in each dataset and the total number of instances in the training dataset is shown below.

Dataset	**Instances**
RAVDESS Speech	1440
RAVDESS Song	1012
TESS	2800
Total	**5252**

After combining the three datasets following is the class distribution of our dataset.

Emotion	**Instances**	**Percentage**
Neutral	588	11.272
Calm	376	7.159
Happy	776	14.775
Sad	776	14.775
Angry	776	14.775
Fear	776	14.775

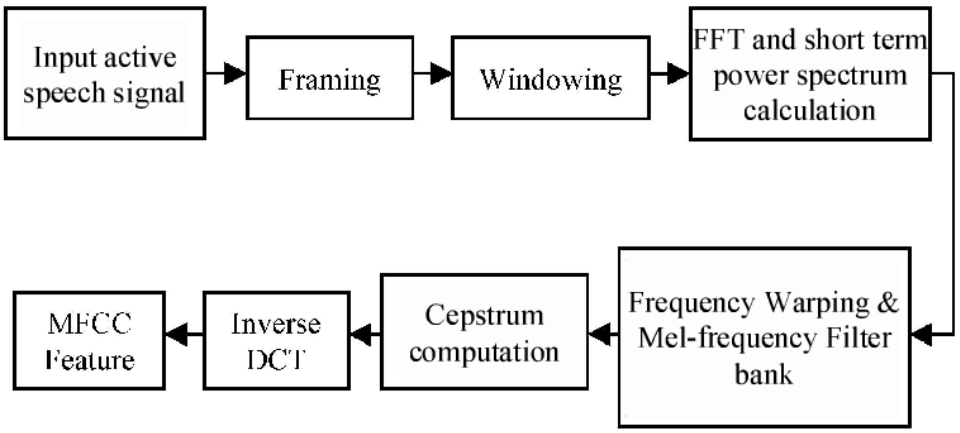
Disgust	592	11.272
Surprised	592	11.272

Here, it is visible that we have fewer instances for calm emotion than for others. But, at an overall level, we have more data points in our dataset for other emotions. This slight class imbalance won't affect our model's performance much. After this, we loaded all sound files into the memory at a sampling rate of 22050 Hz. High sampling rate provided us with data of higher quality but less in size compared to the original one. This helped us in making our model faster.

B. Feature Extraction

In our research, we extract the first 40-order of the MFCC coefficients where the speech signals are sampled at 22050 Hz. The Discrete Cosine Transform (DCT) of the mel log energies were estimated and the first 40 DCT coefficients provided the MFCC values used in the classification process.

The window size used for the feature extraction was 2048 samples combined with a hop length of 512 samples. This helped us in calculating Fast Fourier Transform. In terms of milliseconds, we used a window size of about 93 ms and hop length of 23.22 ms. Hann windowing function was used for avoiding spectral leakage along with overlapping frames. The entire MFCCs extraction pipeline is shown below.



The feature extraction at this minute level helped us in extracting finer details present in the speaker's voice through which we were able to characterize each emotion. After this process, we received a 2D - matrix of size number of coefficients X number of frames for each audio file. But this had a problem of its own. All the files were not of same length clipping the files would have resulted in the loss of data.

Therefore, we transposed these matrices and took a mean along each Mel-frequency Cepstral Coefficient. This provides us with a vector of length 40 for each of the instances. After this, we split our dataset into train, validation and test set in a ratio of 70:15:15. The number of instances after the split in each set is shown in the figure below.

Dataset	**Instances**
Train	3676
Validation	788
Test	788

C. Model Architecture

The model is constructed based on 1D CNN with a fully-connected layer. Each 1D convolutional layer is followed by a ReLU activation layer, a dropout layer and a 1D Max Pooling layer. ReLU activation function is applied to contribute to the sparsity of the network which reduces the interdependence of parameters and alleviates the occurrence of the over-fitting problem. As shown in the table below, the model has three 1D convolutional layers. After the 3rd 1D convolutional layer, the feature maps are flattened into a one-column matrix to fit them into the fully connected layers. The flattened output is then fed to a dense layer then finally softmax function is applied for classification in the last layer. In summary, the whole network is made up of three 1D convolutional layers and

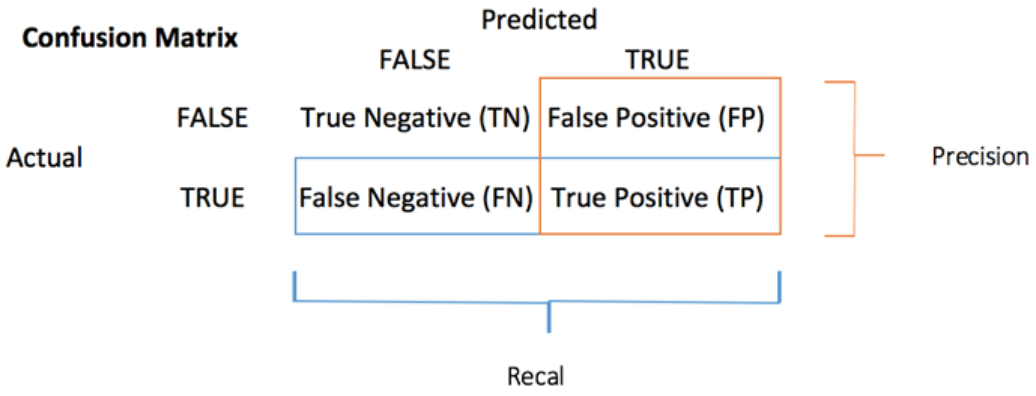
three dropout layers and a dense layer. The model architecture is shown in the table below.

Layer	Input Shape	Output Shape
Conv 1D	40 x 1	40 x 64
ReLU	40 x 64	40 x 64
Dropout	40 x 64	40 x 64
Max Pooling 1D	40 x 64	10 x 64
Conv 1D	10 x 64	10 x 128
ReLU	10 x 128	10 x 128
Dropout	10 x 128	10 x 128
Max Pooling 1D	10 x 128	2 x 128
Conv 1D	2 x 128	2 x 256
ReLU	2 x 256	2 x 256
Dropout	2 x 256	2 x 256
Flatten	2 x 256	512
Dense	512	8
Softmax Activation	8	8 (Output)

6. Evaluation Measures:

6.1 CONFUSION MATRIX

A confusion matrix is a performance measurement method for the evaluation of models in classification problems. It's a kind of table that allows you to understand the performance of the models on a test dataset for which true values are known. The word confusion matrix itself is rather plain, but its associated terms can be a little complicated. Here a simple explanation is given for this technique.



Confusion Matrix is a useful machine learning tool that helps you to calculate Memory, Precision, Accuracy, and F1-Score. It also helps in predicting various mathematical terms such as **Accuracy, True Positive Rate(TPR), True Negative Rate(TNR), False Positive Rate(FPR), False Negative Rate(FNR), and Precision**. All this helps us in understanding machine learning more accurately and to improve it accordingly.

Following are the formulas to calculate some mathematical terms.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where, TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

Precision tells us how many of the correctly predicted cases actually turned out to be positive.

$$Precision = \frac{TP}{TP + FP}$$

Recall tells us how many of the actual positive cases we were able to predict correctly with our model.

$$Recall = \frac{TP}{TP + FN}$$

F1-score is a harmonic mean of Precision and Recall, and so it gives a combined idea about these two metrics. It is maximum when Precision is equal to Recall.

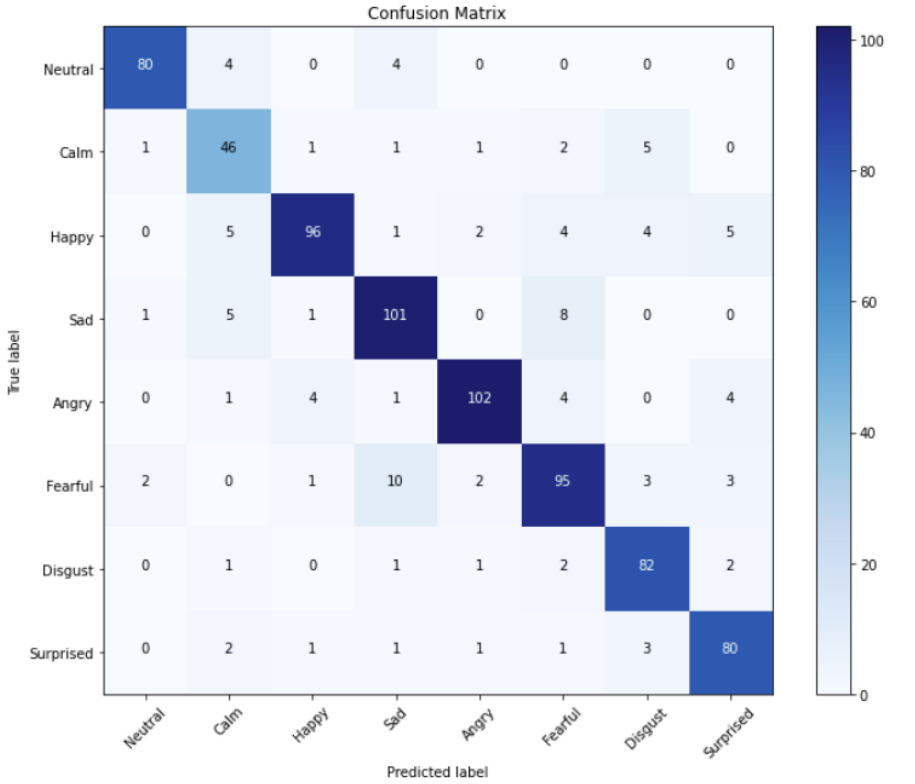
$$F1 - Score = \frac{2 * Recall * Precision}{Recall + Precision}$$

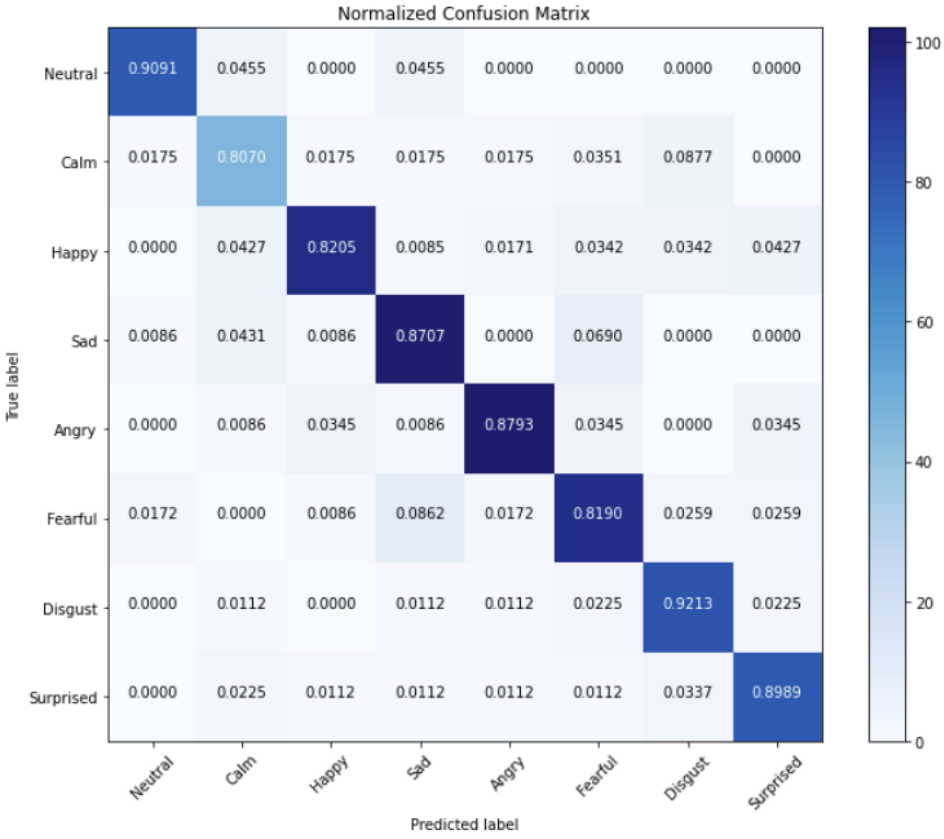
7. Results:

In this section, we'll take a look at the results of this research work. We achieved 86.55% of accuracy and 86.62% of weighted F1-score during the evaluation on the test set. The precision, recall and F1-score for individual classes are shown in the table below.

Emotion	**Precision**	**Recall**	**F1-Score**	**Support**
Neutral	0.952381	0.909091	0.930233	88
Calm	0.718750	0.807018	0.760331	57
Happy	0.923077	0.820513	0.868778	117
Sad	0.841667	0.870690	0.855932	116
Angry	0.935780	0.879310	0.906667	116
Fearful	0.818966	0.818966	0.818966	116
Disgust	0.845361	0.921348	0.881720	89
Surprise	0.851064	0.898876	0.874317	89
Weighted Average	0.869218	0.865482	0.866238	788

Following figures show the non-normalized and normalized confusion matrix for the predictions on the test set.


Confusion Matrix
Neutral
Calm
Happy
Sad
Angry
Fearful
Disgust
Surprised
True label
Predicted label
80 4 0 4 0 0 0 0
1 46 1 1 1 2 5 0
0 5 96 1 2 4 4 5
1 5 1 101 0 8 0 0
0 1 4 1 102 4 0 4
2 0 1 10 2 95 3 3
0 1 0 1 1 2 82 2
0 2 1 1 1 1 3 80
100
80
60
40
20
0


Normalized Confusion Matrix
Neutral
Calm
Happy
Sad
Angry
Fearful
Disgust
Surprised
True label
Predicted label
0.9091 0.0455 0.0000 0.0455 0.0000 0.0000 0.0000 0.0000
0.0175 0.8070 0.0175 0.0175 0.0175 0.0351 0.0877 0.0000
0.0000 0.0427 0.8205 0.0085 0.0171 0.0342 0.0342 0.0427
0.0086 0.0431 0.0086 0.8707 0.0000 0.0690 0.0000 0.0000
0.0000 0.0086 0.0345 0.0086 0.8793 0.0345 0.0000 0.0345
0.0172 0.0000 0.0086 0.0862 0.0172 0.8190 0.0259 0.0259
0.0000 0.0112 0.0000 0.0112 0.0112 0.0225 0.9213 0.0225
0.0000 0.0225 0.0112 0.0112 0.0112 0.0112 0.0337 0.8989
100
80
60
40
20
0

8. Conclusion and Future Scope:

This paper proposes a model based on the 1D convolutional network for speech emotion recognition. Features were mainly extracted by complementary features, then based on 1D CNN to capture emotional features. We combined three different datasets and trained a model over it. This enabled us to create a more generalized model which can identify the emotions of the speaker in different settings as well. Through the results of this research, it is clear that our proposed method works efficiently in speech emotion and recognition with good performance.

Nevertheless, we think that more research on this topic can be done. Here, we trained the model completely from scratch for speech emotion recognition. But, we believe that there is still a scope of improvement in performance if we go towards modern transfer learning and ensemble learning methods.

9. References:

1. I.T. Kun Han, D. Yu, Speech emotion recognition using deep neural network and extreme learning machine, Interspeech (2014) 223–227.
2. A.M. Badshah, J. Ahmad, N. Rahim, S.W. Baik, Speech emotion recognition from spectrograms with deep convolutional neural network, in: 2017 International Conference on Platform Technology and Service (PlatCon), IEEE, 2017, pp. 1–5.
3. S. Mittal, S. Agarwal, M.J. Nigam, Real-time multiple face recognition: a deep learning approach, in: Proceedings of the 2018 International Conference on Digital Medicine and Image Processing, ACM, 2018, pp. 70–76.
4. H.-S. Bae, H.-J. Lee, S.-G. Lee, Voice recognition based on adaptive mfcc and deep learning, in: 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), IEEE, 2016, pp. 1542–1546.
5. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016) 770–778.

6. K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, Y.-H. Chen, Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds, in: ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 5866– 5870.
7. A. Tawari and M. Trivedi, "Speech-based emotion classification framework for driver assistance system," in 2010 IEEE Intelligent Vehicles Symposium, San Diego, CA, USA, 2010, pp. 174-178.
8. H. Alipour, D. Zeng and D. C. Derrick, "AdaBoost-based sensor fusion for credibility assessment," in 2012 IEEE International Conference on Intelligence and Security Informatics, Arlington, VA, USA, 2012, pp. 224-226.
9. S. Tokuno, G. Tsumatori, S. Shono, E. Takei, T. Yamamoto, G. Suzuki, S. Mituyoshi and M. Shimura, "Usage of emotion recognition in military health care," in 2011 Defense Science Research Conference and Expo (DSR), Singapore, 2011, pp. 1-5.
10. Neiberg, D.; Elenius, K.; Karlsson, I.; Laskowski, K.: Emotion recognition in spontaneous speech. In: Proceedings of Fonetik, pp. 101–104 (2006).
11. Chenchah, Farah, and Zied Lachiri. "Acoustic emotion recognition using linear and nonlinear cepstral coefficients." International Journal of Advanced Computer Science and Applications 6.11 (2015): 135-138.
12. Srinivas Parthasarathy and Ivan Tashev (2018), Convolutional Neural Network Techniques For Speech Emotion Recognition, Microsoft Research
13. N. B. Wunarso and Y. E. Soelistio, “Towards Indonesian speech-emotion automatic recognition (i-spear),” in 2017 4th International Conference on New Media Studies (CONMEDIA), pp. 98–101, Nov 2017.
14. A. A. A. Zamil, S. Hasan, S. M. Jannatul Baki, J. M. Adam, and I. Zaman, “Emotion detection from speech signals using voting mechanism on classified frames,” in the 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), pp. 281–285, Jan 2019.
15. C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, Iemocap: interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (2008) 335.

16. S. Yoon, S. Byun, K. Jung, Multimodal speech emotion recognition using audio and text, in: 2018 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2018, pp. 112–118.
17. LIVINGSTONE, S. R., AND RUSSO, F. A. The Ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one 13, 5 (2018), e0196391.
18. Pichora-Fuller, M. Kathleen, and Kate Dupuis. Toronto Emotional Speech Set (TESS). Scholars Portal Dataverse, 2020. DOI.org (Datacite), doi:10.5683/SP2/E8H2MF.
19. J. Lyons, “Mel frequency cepstral coefficient (MFCC) tutorial,” Practical Cryptography, 2015.