Speech Recognition with Deep Learning

Speech Recognition with Deep Learning

[1]Lokesh Khurana, [2]Arun Chauhan, [3]Dr. Mohd Naved, [4,*]Prabhishek Singh

[1, 2, 4]Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, India

[3]Assistant Professor, Jagannath University

[1]lokeshkhurana98@gmail.com

[2]arunchauhan414@gmail.com

[3]mohdnaved@gmail.com

[4]prabhisheksingh88@gmail.com

*Corresponding author email: prabhisheksingh88@gmail.com

Abstract

The human voices are very flexible and contains a mess of sentiments or emotions. Feeling or emotions in speech incorporates extra vision about human activities. Recognition of various emotions from the human speech signal is very stretching ingredient in human computer interaction. Through in addition analysis, we can higher recognize the rationale of human beings or people, whether they are not happy with the service clients, happy customers, encouraging folks or inspiring fans. Deep Learning strategies have been as of lately proposed as an option in contrast to conventional techniques in Speech Emotion Recognition (SER). The Emotion of a speaker can be easily govern by the humans because it the human nature to understand the complexion of a person by just guessing the flow of speech, but the domain of emotion or sentiment recognition in the course of machine learning is an open circle of research. In this intended project, we perform speech information evaluation on speaker discriminated speech signals to hitupon the warmth of the person speakers involved inside the verbal exchange. we're analyzing exceptional techniques to perform speaker discrimination and speech evaluation to locate efficient algorithms to carry out this task.

Keywords— SER; emotions; features; algorithms; extraction

1. Introduction

In spite of the fact that emotion recognition from discourse is a for the most part new field of research, it has various potential applications. In human-human or human-machine affiliation framework, emotion detection frameworks could outfit customers with improved organizations by being flexible to their sentiments. In virtual universes, emotional acknowledgment could help reproduce progressively reasonable symbol cooperation.The group of work on distinguishing feeling in discourse is very restricted. At present, analysts are as yet discussing what highlights impact the recognition of feeling in discourse. There is additionally significant vulnerability with respect to a better distinguishing emotion algorithm, and which emotions to group conjointly. For a machine to comprehend the attitude/mind-set of the people through a discussion, it has to realise who are associating in the discussion and what is spoken, so we execute a speaker and discourse recognition framework or system first and perform speech examination on the information separated from earlier procedures [22-24]. Understanding the temperament of people can be helpful in numerous examples. For instance, computers and machines that has the capacity to see and react to human non-lexical correspondence, for example, feelings. In such a case, in the

wake of distinguishing humans' emotions, the machine could alter the settings concurring his/her needs and inclinations.

2. Methodology

2.1 Dataset for Speech Signals

RAVDESS: The RAVDESS database is gender adjusted comprising of 24 professional and expert actors of which 12 are male and 12 are female actors. Speech portion of dataset incorporates 1500 audio files as input. The audio file consist of 8 different emotion short audios in it. These emotional audio samples are in wave format. Speech emotions dataset are as follow disgust, happy, angry, fearful, sad, surprise, and calm expressions and dataset also contains songs with emotions i.e. calm, happy, sad, angry, and fearful emotions.

SAVEE: The SAVEE dataset comprises of accounts from 4 male on-screen characters in 7 distinct feelings, 480 British English articulations altogether. 'DC', 'JE', 'JK' and 'KL' are the four male speakers recorded in the SAVEE database. There are 15 sentences for every one of the 7 feeling classifications, and one document for each sentence. It incorporates 'outrage', 'disturb', 'dread', 'joy', 'neutral', 'pity' and 'surprise'. I made redid dataset by utilizing this two dataset. We have less examples of 'calm' feeling. Henceforth I discarded 'calm' tests to adjust the dataset. Our dataset contain 7 folders, each represents to the diverse emotion. Contain Separate emotion's voice/discourse in each different folder.

Librosa library of Python is used in order to process and extract the features from the files “Audio Files”. Librosa is a library used for speech, audio and music analysis.It gives the structure squares important to make audio data recovery system. Utilizing the librosa library, Option to extricate features i.e MFCC(Mel Frequency Cepstral Coefficient) is also possible. MFCCs are a component generally utilized for the automatically recognition of the audio and the speaker. We additionally separated the females and males voice by the utilizing the identifiers. Separation of female and male voices increased the accuracy by 15%. It could be a direct result of the pitch of the voice was influencing the outcomes. Every sound record gave us numerous features which were fundamentally exhibit of numerous qualities [25-28].

Here are the waveforms for Angry scream, Normal Speech and laughter in the figure below.

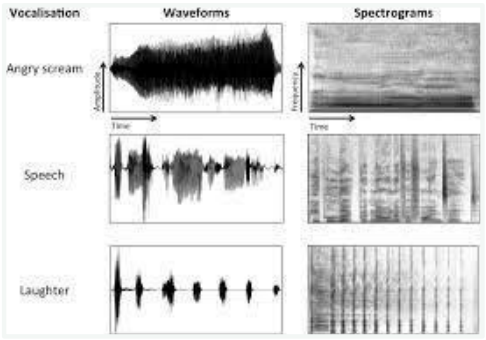

Fig.1: Waveform and Spectrograms for laugh, anger and normal speech [16]

2.2 Feature Extraction

So as to extricate the features from the discourse or speech signals, Mel Frequency Cepstral Coefficients (MFCC) must be extricated from the signals. The time-space portrayal of sound is astoundingly unpredictable, and in its extraordinary structure, it doesn't give magnificent information into key properties of the signals. Considering this nature of sound signals, we map this time-space portrayal into moreover telling features. The most immediate strategy incorporates choosing the ordinary essentialness of the signals which quantifies the "volume" of the speaker. At the point when a sound signal shows up it contains different furthermore bits of information into emotions, similar to the most extreme, mean, standard

deviation, least, and range of both the signals and the ranges. These may demonstrate changes in the volume or contribute that can be valuable deciding feeling.

For both the signals and the spectrums, likewise determine the skewness, “the measure of departure of horizontal symmetry in the signal”, and the kurtosis, “the measure of height and sharpness of central peak, relative to a standard bell curve

2.3 Windowing of waveform

The Mel-frequency Cepstrum catches qualities of the various frequency or recurrence of the signal constitute on the Mel scale, which intently adjusts to the non-linear nature of human beings for hearing. Mel-frequency Cepstrum Coefficient (MFCC) illustrate the “spectrum of the spectrum”. The power of frequency spectrum is mapped onto the Mel scale and afterward log of frequency spectrum is taken in order to derive the MFCC.

Changes in the pitch are measured over time with the help of two scales which are “coarse time scale” and “fine time scale”. For coarse estimation, the signal is separated into 3 sections (starting, center, and end), and the piece of the signal with the most noteworthy average pitch is utilized to decide if the pitch rises or falls after some course of time. For fine estimation, the prevailing frequency of each windowed test is contrasted with the predominant frequencies of the windowed tests promptly preceding and following. This distinction is recorded in the component or feature vector.

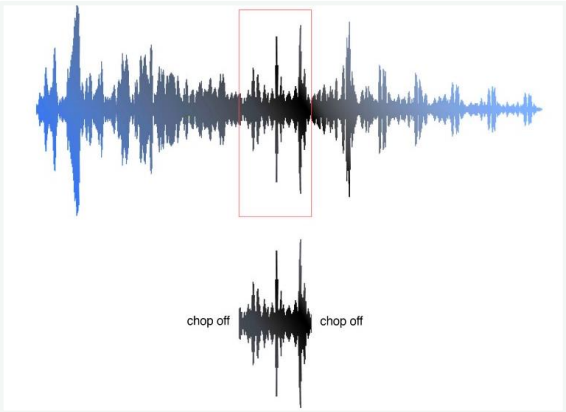

Fig.2 Signal Spectrum [17]

3. Models

3.1 CNN (Convolutional Neural Network)

It is a “deep learning algorithm” which is capable to take speech/image as input and then assign the weights and biases to the objects present in the speech/image. It is the net which uses or requires the technical operation of convolution to search for a particular patterns or figures. It has flashlight structure. Each flashlight represents a neuron in the Convolutional Neural Network (CNN).

The input, speech/image are passed to the very first layer of Convolutional layer. The output is obtained as the activated map from the convolutional layer. There are filters present in the convolutional layer which are responsible for the extraction of the features from input and pass them further. Every filter will give an alternate feature to help for the prediction of right class. Padding (zero padding) is used to hold or retain the actual size of the image. legitimate padding technique is used as it will help to reduce the count of features for better performance of the model.The Pooling layer is used for the reduction of the dimensionality. Because of the pooling layer net will possibly focuses on only the most relevant patterns discovered by the convolutional layer and RELU layer [29-31].

Before the prediction made, a few number of convolutional layers and pooling layers are figured. The primary errand of convolutional layer is to separate the most explicit features from the dataset. These specific features will extracted for the model by going deeper in the network.

3.2 LSTM Networks

Usually Long Short Term Machine is just called LSTM. LSTM are a kind of Recurrent Neural Network, they are best fit for the problems related to sequences prediction.

LSTM are most required for the complex domain of problems like speech recognition or machine translation. LSTM is one of the complicated domain of deep learning. LSTM, on the direct intended to overcome the long-term problem of dependency.

All the recurrent neural networks have a kind of a chain of modules which are repeated. The signals will only flow in one direction, so the signals have unidirectional flow from input to the output which is considered as one layer at a time. This deep learning model has a basic structure with the implicit feedback loop.

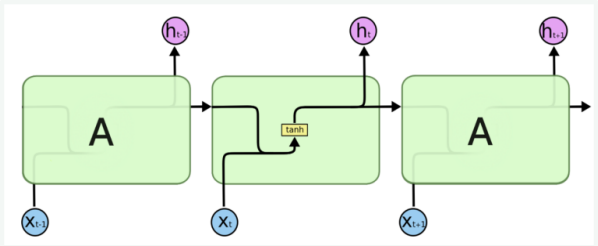

Fig.3 LSTM Model Architecture [18]

4. Results

The testing and training loss in the dataset is shown with the comparison between three neural network technique to present the best network for the speech emotion recognition. The graphical representation shows the error in the training and testing of the dataset getting reduced as the number of epoch (hyperparameter) for the training model get increased.

1 Epoch = 1 Forward pass + 1 Backward pass for **ALL** training samples. Batch Size = Number of training samples in 1 Forward pass and 1 Backward pass. (With increase in Batch size, required memory space increases.) Number of iterations = Number of passes i.e. 1 Pass = 1 Forward pass + 1 Backward pass (Forward pass and Backward pass are not counted differently.)

4.1 For CNN

The convolutional neural network (CNN) model had the training accuracy of around 47% with 13 layers, "tan h" activation function, batch size of around 350 epochs.

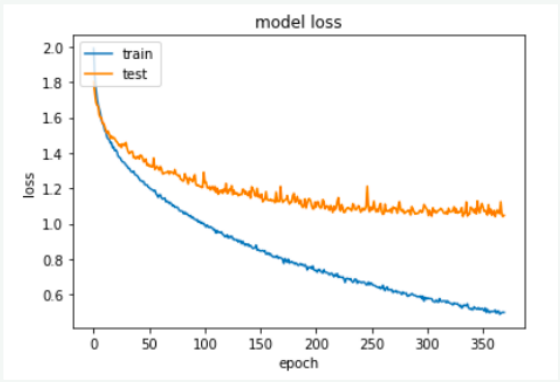

Fig 4. Training/Testing [19]

4.2 For MLP

The multi-layer perceptron (MLP) model had the training accuracy of around 25% with 8 layers, batch size of around 550 epochs.

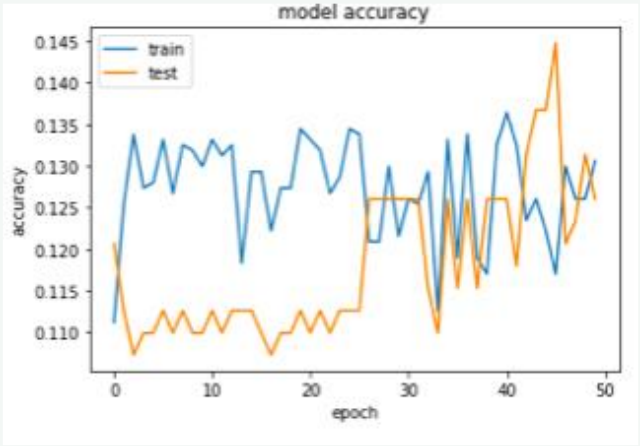

Fig. 5: MLP Training/Testing [20]

4.3 For LSTM

The long-short term machine (LSTM) model had the lowest training accuracy of around 15% with 5 layers, batch size of around 50 epochs.

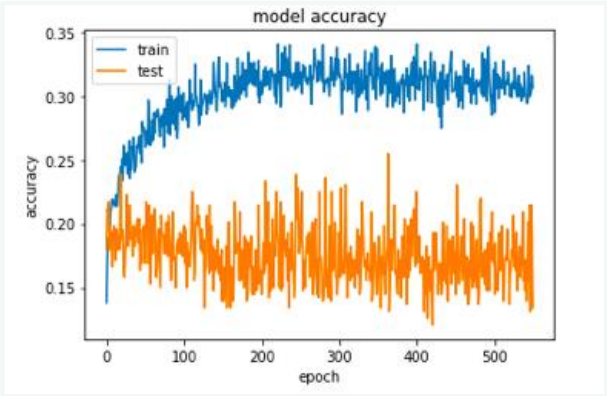

Fig. 6: LSTM Training/Testing [21]

5. Conclusion

This Project gives the machine an ability to have a proper conversation between machine and the humans. With the ability of emotion recognition from speech the machine can understand the human emotions through human voice with better accuracy and reply them with proper ease and have a seamless conversation like human beings.

Reference

[1] https://medium.com/@jonathan_hui/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9

[2] https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

[3] IEEE Research Paper: https://ieeexplore.ieee.org/abstract/document/8805181

[4] https://machinelearningmastery.com/convolutional-layers-for-deep-learning-neural-networks/

[5] M.M.H.E. Ayadi, M.S. Kamel, F. Karray, "Survey on speech emotion recognition: Features classification schemes and databases", *Pattern Recognition*, pp. 572-587, 2011.

[6] https://www.intechopen.com/books/social-media-and-machine-learning/automatic-speech-emotion-recognition-using-machine-learning

[7] https://www.sciencedirect.com/science/article/abs/pii/S1746809420300501

[8] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture," Acoust. Speech, Signal Process., vol. 1, pp. 577–580, 2004.

[9] C.-W. Huang and S. S. Narayanan, "Characterizing Types of Convolution in Deep Convolutional Recurrent Neural Networks for Robust Speech Emotion Recognition," pp. 1–19, 2017.

[10] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture," Acoust. Speech, Signal Process., vol. 1, pp. 577–580, 2004.

[11] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," 2017 Int. Conf. Platf. Technol. Serv., pp. 1–5, 2017.

[12] J. Schmidhuber, “Deep Learning in neural networks: An overview,” Neural Networks, vol. 61, pp. 85–117, 2015.

[13] https://www.ijrte.org/wp-content/uploads/papers/v7i4s/E1917017519.pdf

[14] F. Dipl and T. Vogt, “Real-time automatic emotion recognition from speech,” 2010.

[15] H. M. Fayek, M. Lech, and L. Cavedon, “Evaluating deep learning architectures for Speech Emotion Recognition,” Neural Networks, vol. 92, pp. 60–68, 2017.

[16] F. Dipl and T. Vogt, “Real-time automatic emotion recognition from speech,” 2010.

[17]https://towardsdatascience.com/ok-google-how-to-do-speech-recognition-f77b5d7cbe0b?gi=ba575dcaaeae[16] F. Dipl and T. Vogt, “Real-time automatic emotion recognition from speech,” 2010.

[18]https://www.programmersought.com/article/4093817767/

[19]https://github.com/MITESHPUTHRANNEU/Speech-Emotion-Analyzer/tree/master/images

[20]https://github.com/MITESHPUTHRANNEU/Speech-Emotion-Analyzer/tree/master/images

[21]https://github.com/MITESHPUTHRANNEU/Speech-Emotion-Analyzer/tree/master/images

[22] Singh, Prabhishek, and Raj Shree. "Statistical modelling of log transformed speckled image." International Journal of Computer Science and Information Security 14.8 (2016): 426.

[23] Singh, Prabhishek, and Raj Shree. "Quantitative Dual Nature Analysis of Mean Square Error in SAR Image Despeckling." International Journal on Computer Science and Engineering (IJCSE) 9.11 (2017): 619-622.

[24] Diwakar, Manoj, and Prabhishek Singh. "CT image denoising using multivariate model and its method noise thresholding in non-subsampled shearlet domain." Biomedical Signal Processing and Control 57 (2020): 101754.

[25] Singh, Prabhishek, and Raj Shree. "A New Computationally Improved Homomorphic Despeckling Technique of SAR Images." International Journal of Advanced Research in Computer Science 8.3 (2017).

[26] Diwakar, Manoj, et al. "Latest trends on heart disease prediction using machine learning and image fusion." Materials Today: Proceedings (2020).

[27] Dhaundiyal, Rashmi, et al. "Clustering based Multi-modality Medical Image Fusion." Journal of Physics: Conference Series. Vol. 1478. No. 1. IOP Publishing, 2020.

[28] Kumar, Neeraj, et al. "Flood risk finder for IoT based mechanism using fuzzy logic." Materials Today: Proceedings (2020).

[29] Jindal, Muskan, et al. "A novel multi-focus image fusion paradigm: A hybrid approach." Materials Today: Proceedings (2020).

[30] Diwakar, Manoj, et al. "A comparative review: Medical image fusion using SWT and DWT." Materials Today: Proceedings (2020).

[31] Maurya, Awadhesh Kumar, Ajay Kumar, and Neeraj Kumar. "Improved chain based cooperative routing protocol in wsn." Journal of Physics Conference Series. Vol. 1478. No. 1. 2020.