International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.133

# Speech Emotion Recognition Using LSTM Model

Y H SaiDhruv
*Department of Data Science and Business Systems, School of Computing, College of Engineering and Technology,*
*SRM Institute of Science and Technology,*
Kattankulathur, Chennai, 603203, India
ys5544@smrist.edu.in

Priyadarsini k
*Department of Data Science and Business Systems, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology,*
Kattankulathur, Chennai, 603203, India
priyadak@srmist.edu.in

M Vishnu Vardhan
*Department of Data Science and Business Systems, School of Computing, College of Engineering and Technology,*
*SRM Institute of Science and Technology,*
Kattankulathur, Chennai, 603203, India
mv1614@smrist.edu.in

Jeba Sonia J
*Department of Data Science and Business Systems, School of Computing, College of Engineering and Technology,*
*SRM Institute of Science and Technology,*
Kattankulathur, Chennai, 603203, India
jebas@srmist.edu.in

***Abstract*—This approach of speech emotion recognition(SER) is a problem in the subject of deep learning. The primary objective of this study is to identify a variety of feelings based on the input of audio utterances as well as the Rnn networking. We used the long short-term memory networks (LSTM) model because it deals with speech features and aims to improve speech recognition rates. Initially, this method of speech emotion was done by machine learning approach but recently deep learning techniques are proven to be an alternative to SER providing better accuracy.We used the dataset's audio files as input, performed feature extraction, and trained the model. Considering researching the published works from prior periods' research and putting the suggested models through their paces on our own datasets, We expect that the overall file classification efficiency will be 99.73**

***Index Terms*—Recurrent Neural Network(RNN),Long Short- term memory(LSTM), Speech emotion recognition(SER), Deep learning.**

## I. INTRODUCTION

Making it feasible for human-machine contact is the main goal of this mission speech emotion recognition (SER) utilising deep learning method. Speech is best way of interaction between humans. So, this is very exciting and interesting for researchers to find ways to establish communication between machines and humans. This topic of SER is quite tough because there are many factors that might effect the raw input sound signals such as age, gender, noise and etc. Nowadays, the majority of our focus is directed into deep learning since, within a relatively short amount of time, it is providing us with improved results in every area, particularly SER. Deep learning makes it possible to use sophisticated models and learn various data representations. Deep learning's main disadvantage is that it needs a large data in order to perform better than any other approaches. Firstly, a machine learning technique was used to perform this SER technique; however, once a large dataset was used, this machine learning method produced superior results. Despite the fact that speech technology has advanced significantly SER in previous years, but a bettersystem is still required. Recognition of human sentiments via machines, which might further improve the relationship between humans and machines [1][4]. The primary and most direct method of information transmission is speech. It is able to express a vast amount of emotional information by means of the feelings that it feels and the manner in which it displays those feelings in response to items, scenarios, or happenings. This ability allows it to communicate a wide range of sentiments. Its database contains a vast amount of information covering a variety of topics. In recent years, a significant number of studies and research initiatives have focused their attention on developing methods for the automated recognition of individual emotions through the study of human voices and facial expressions. These methods have been developed in an effort to improve the accuracy with which computers can read face expressions and voices. There is a significant rise in the number of studies pertain to this topic as a direct result of the fact that automated emotion detection systems may be utilised for a variety of purposes within a variety of settings. This is a direct result of the fact that automated emotion detection systems may be used in a variety of settings.Examples of the applications for these studies and their planned use include the following systems: Education: A course system for online learning can identify disinterested students so that the style or difficulty of the material can be changed. It can also offer emo- tional bribes or concessions. Automobile: The driver's internal condition and driving performance are frequently connected. As a result, these systems can be employed to enhance driving performance and the driving experience. Security:It's possible that intense emotions like fear and anxiety may be used as support systems in public settings if we acknowledged them and acknowledged their significance. Communication: When an interactive voice response system and an automatic emotion detection system are combined in call centers, customer service may be enhanced. Health: Individuals who suffer from autism may find it easier to adjust their social behaviour if they have access to portable technology that helps them understand their own sensations and feelings [5].

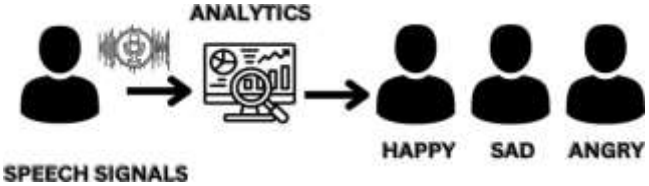


Fig. 1. Process of SER

DOI: rp-9788770040723.133

## II. Related works

An intriguing piece of research that Pan produced made use of MFCC and support vector machines as features. With regard to his Chinese dataset as well as EMoDB[6], he achieved astounding accuracies of 91.3043percentage and 95.087 percentage, respectively. The Arabic television programme ”The Exact reverse Path,” according to Ashraf Khalil in [6], served as the source for his set of samples. He used SVM and got a classification rate of 77 percentage.[3][7] We are aware that Yixiong Pan in [8] used support vector machine to classify emotions into 3 groups with 95.1percentage correctness using datasets from the Berlin Database of Emotional Speech [1][8]. The finest outcome in this field has thus far been published, and it is this one. In [9], S. Lalitha uses a support vector machine (SVM) classifier in conjunction with pitch and prosody characteristics to achieve 81percentage accuracy across 7 classes from the full Berliner Dataset of Emotional Speech. In [10], Yu Zhou merged spectral and prosodic features using the gaussian mixture model superior vector- based SVM, and he showed 88.3 percentage efficiency on five classes of the Chinese-LDC corpus. In [11], Fei Wang used a mixture of various features,Deep Auto Encoder, and SVM to achieve an accuracy of 83.7 percent on six classes from the Chinese emotion corpus CASIA. [1] SVM is employed in some other Arabic dataset created by Al-faham and Ghneim. [12]. where they hit their target 93.1 percent of the time HabibaDahmani proposed in [13][3] a fully automated system for voice recognition for the Algerian language. She used data from the Algerian television show ”Red Line” for her research. The total number of sound samples is 1443. Among the several classification strategies used were KNN, Logistic regression, and Random Forests. Their F-value came out to be 0.48. [3] In [14][1], JianweiNiu reported 92.3percentage accurate on six classes of 7676 speaking Mandarin Chinese phrases employing a range of characteristics in their identification system and integrating deep beural networks and Hidden Markov Mode.

H.M. Fayek in [15] studied multiple DNN designs using two independent databases, eNTERFACE [16] and SAVEE [17][1], with 6 and 7 categories, respectively.[1] John Kim introduced a framework called EmNet in that includes extraction offeatures, feature normalisation, 4 CNN layers, and 2 LSTM layers.[18]. EmNet performed at a rate of 88.9 percentage when tested on the EMO-DB dataset.[3] Yasser Hifny and Ahmed Ali [19] developed two neural architectures to address the problem of voice emotion detection. The first architecture employed the CNN-LSTM-DNN model, whereas the second design was based on the CNN model. On the KSUEmotions dataset [20], the accuracy of the first model was 87.2 percent, whereas that of the second model was 85Noroozi et al. Based on the evaluation of both visual and aural data, proposed a customisable approach for recognising emotions. In his study, 88 traits—Mel Frequency Cepstral Coefficients [MFCC] and filter bank energies [FBEs]—were employed to cut down the dimension of feature extraction previously performed using the Principal Component Analysis. Bandela et al. coupled the Teager Energy Operator (TEO) as In his study, 88 traits—Mel Frequency Cepstral Coefficients [MFCC] and filter bank energies [FBEs]—were employed to cut down the dimension of feature extraction previously performed using the Principal Component Analysis.a prosodic feature with the auditory feature known as the MFCC in order to detect five emotions using the Berlin Emotional Speech collection and the Gaussian Mixture Model classifier .The 13 MFCC generated from voice files were utilised by Zamil et al. as spectral characteristics, in their proposed model to classify the seven sentiments using the Logistic Model Tree method.[21]

## III. METHODOLOGY

### A. *Data*

In order to evoke each of the seven emotions, two actors (ages 26 and 64) recite a selection of 200 target words in the carrier phrase ”Speak the word” (anger, fear, disgust, pleasant surprise, sadness, happiness and neutral). There are 2800 data points in all (audio files). The data is structured in this manner so that each of the two female actors and their feelings has its own folder. It includes an audio file with the two hundred target words. The audio file format is WAV. After importing the dataset, we saw that it had around 2800 samples. After importing the dataset, we will create a data frame with audio files and labels. Then we do an exploratory analysis of the data to examine if all of the classes, such as fear, anger, disgust, neutral label, sad, ps, and happy, are distributed uniformly. Finally we assign responsibilities to the spectrogram and wave- plot. The waveform of an audio track file is displayed using a waveplot, while the frequency levels are displayed using a spectrogram. We are presently developing a feature extraction method for audio files. Following feature extraction, we will now construct the LSTM model. A thick two dimensional linear layer with hidden units, dropout regularisation to prevent overfitting by removing a portion of the data, the limited categorization To calculate the difference between true and predicted labels, use the cross-entropy method, and the Adam optimizer to automatically update the information gain during training are all part of the neural network model. Now we will train the model, and all of the results will be shown at the conclusion of each epoch of training, and the results willbe plotted, and we will see whether we get better outcomes with this model.

### B. *LSTM model using RNN Architecture*

LSTM is a type of Recurrent Neural Network (RNN) that can develop protracted dependencies. In this research, a type of speech processing approach for LSTM networking structure is given to contend with speech characteristics, with the goal of boosting the speech recognition rate. A LSTM unit conducts the memorizing in this model, the Dropout unit periodically changes the parameters of a chunk of the data to zero to prevent against overfitting, and also the Dense units include convolution layer connected to the degree of freedom the model will have to adapt to the data. The more complicated the data, the more degrees of freedom the model requires, all while avoiding overfitting . If the accuracy of the training and test sets differs (for example, accuracy rate is 98 percent and test accuracy is 88 percent), the data has been overfit. Stop iterating when the validation metric of choice (accuracy in this example)

DOI: rp-9788770040723.133

begins to drop. Using audio data, we may build an RNN by having started with a basic model and gradually adding layers until it can forecast the data to the top of its abilities. To understand where this border is, adapt the structure until your simulation tends to overfit the data, then go back and remove layers. Check for performance differences in between test and training information and use Drop - outs layers to avoid over - fitting to the training data.


Fig. 2. LSTM Model

*C. Experiment*

Our framework was developed using an LSTM model with an RNN architecture.About 50 epochs were applied to the RNN design. Every sampling was successful and included samples from various classes (sadness, fear, disgust, happi- ness,anger, pleasant surprise, and neutral), there were precisely the same amounts of segments from each class in every batch as shown in fig. 3. Because each class included 400 data frames, the total number of data frames for all emotions was 2800. After the exploratory data analysis, We created a waveplot and spectrogram for each class, as shown in Fig 4, Fig 5, which shows the wave plot and spectrogram of fear, Fig 6, Fig 7 showing the wave plot and spectrogram of angry, Fig 8, Fig 9 showing the wave plot and spectrogram of disgust, Fig 10, Fig 11 showing the wave plot and spectrogram of neutral, Fig 12, Fig 13 showing the wave plot and spectrogram of sad,Fig 14, Fig 15 showing the wave plot and spectrogram of ps, Fig 16, Fig 17 showing the wave plot and spectrogram of happy respectively.A sample audio of an emotional speech is also included in each session. The spectrogram's colour pattern demonstrates that lower pitched sounds have darker colours, while higher pitched tones have vibrant colours. Following that, the audio files are subjected to the extracting features approach. With the same file size, the speech length was restricted to three seconds. MFCC features will be recovered with a limit of 40, and the mean will be used as the final feature.The extracted features of the audio files will then be returned in the following phase, and xmfcc will assist in the display of the features derived from the data.The processing time required to display the generated features rises in proportion to the amount of samples in the dataset. After that, the list is transformed to a one-dimensional array, which we receive as (2800,40). After converting the input to a single- dimensional array, we split it (2800,40,1). The shape in the x region represents the number of samples in the dataset and the number of features in a single dimension array. The y-axis contour shows the number of samples and output classes. We will now construct the LSTM model. The loss in this model is called "sparse categorical cross entropy," which computes the cross-entropy loss between true labels and predicted labels, and the learning rate of the model is automatically changed over the number of epochs.

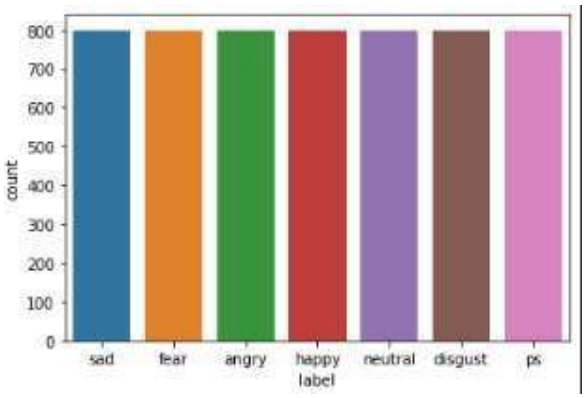


Fig. 3. Exploratory data analysis of our dataset

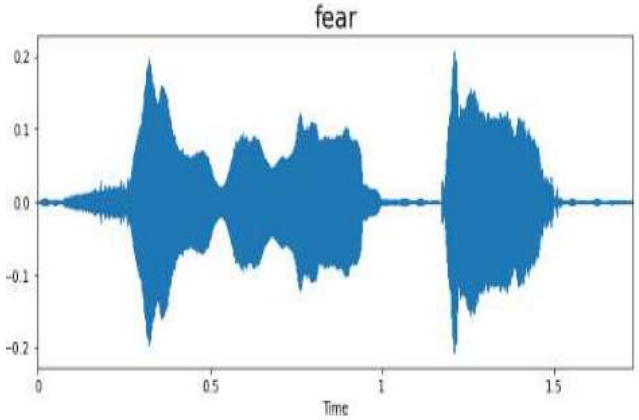


Fig. 4. waveplot of fear emotion

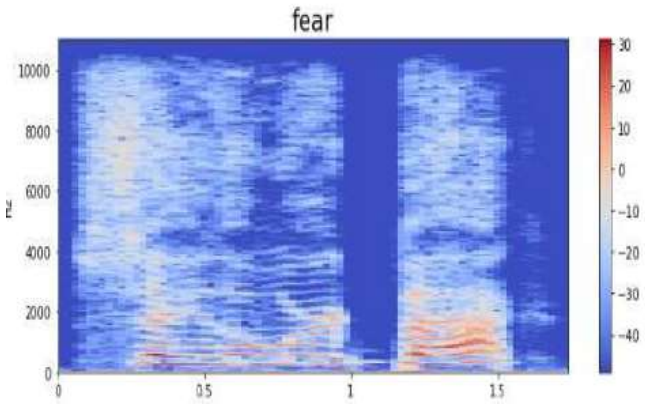


Fig. 5. spectrogram of fear emotion

DOI: rp-9788770040723.133

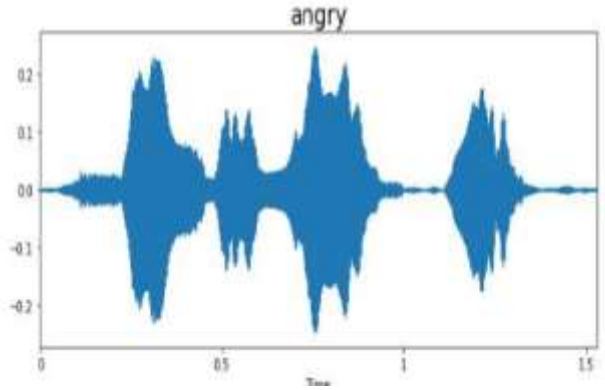


Fig. 6. waveplot of anger emotion

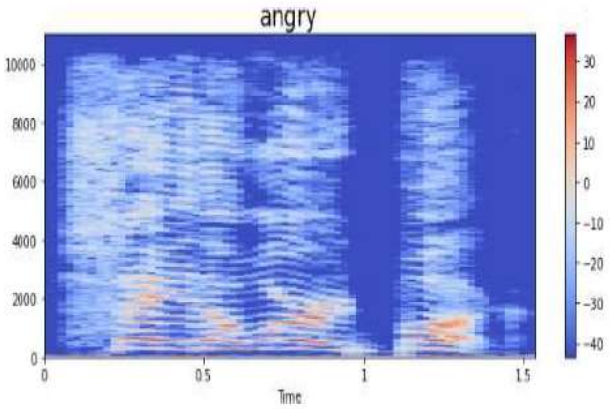


Fig. 7. spectrogram of anger emotion

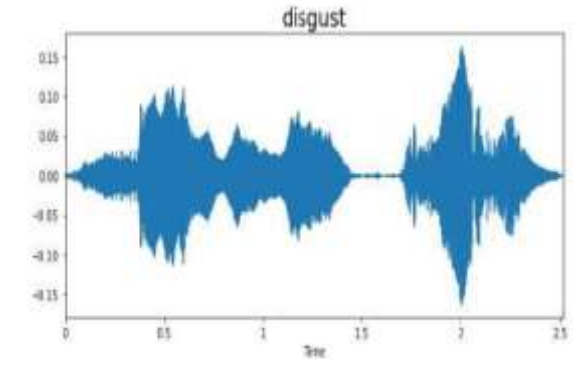


Fig. 8. waveplot of disgust emotion

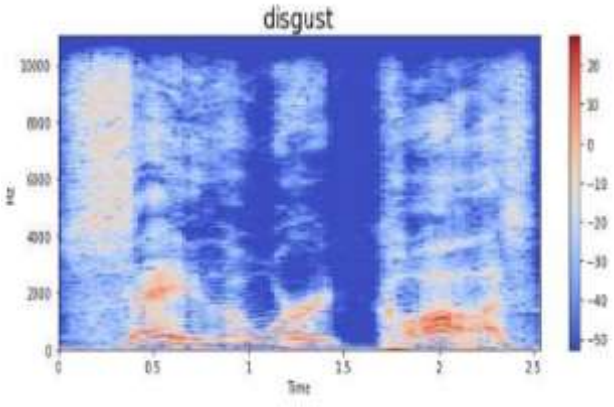


Fig. 9. spectrogram of disgust emotion

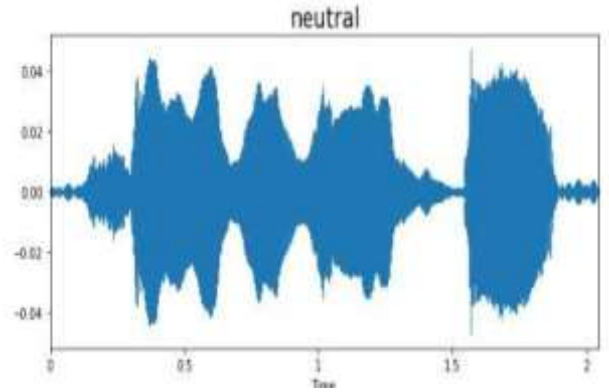


Fig. 10. waveplot of neutral emotion

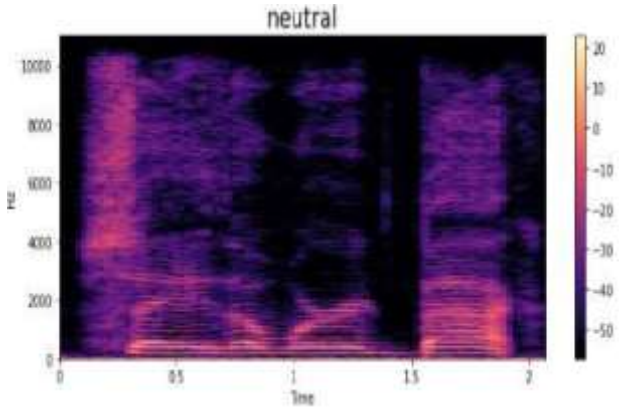


Fig. 11. spectrogram of neutral emotion

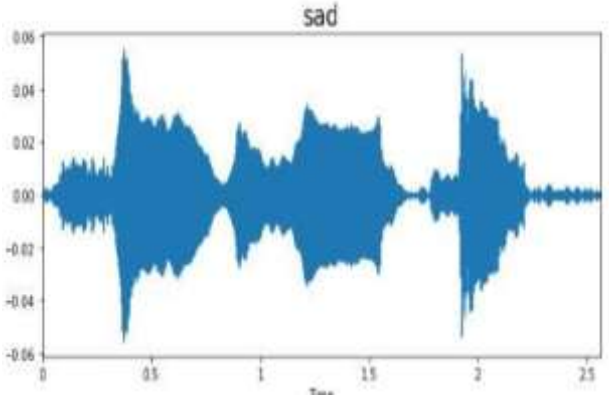


Fig. 12. waveplot of sad emotion

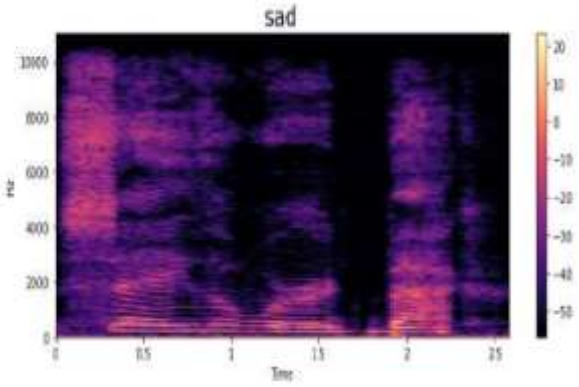


Fig. 13. spectrogram of sad emotion

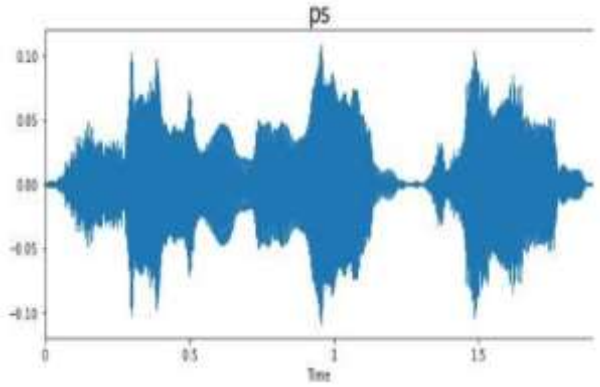


Fig. 14. waveplot of ps emotion

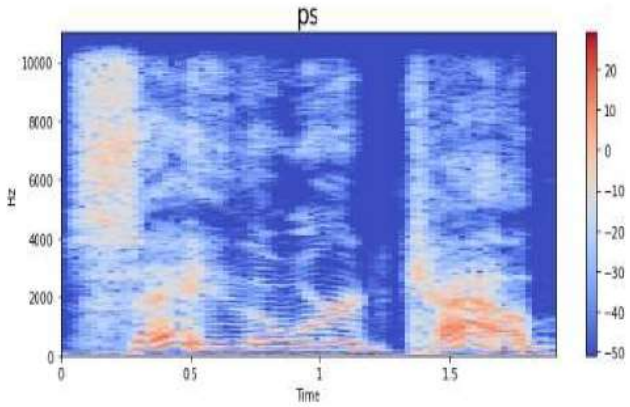


Fig. 15. spectrogram of ps emotion

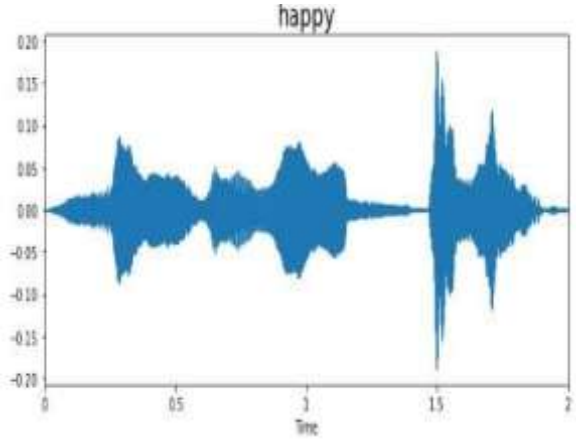


Fig. 16. waveplot of of happy emotion

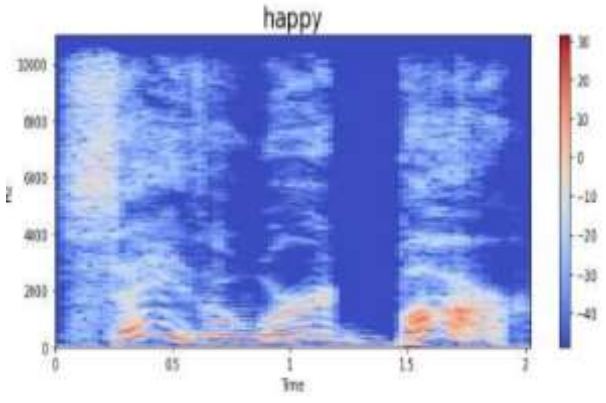


Fig. 17. spectrogram of happy emotion

## IV. RESULTS

Based on the research that we have done,we claim our accuracy to be 99.73 percent. In this model, the memorizing is done by an LSTM unit, the Dropout unit occasionally resets the parameters of a portion of the data to zero to avoid over fitting, and the Dense units also have a convolution layer linked to the amount of flexibility the model will have to adjust to the data. More degrees of freedom are needed in the model as the complexity of the data increases, but over fitting must be avoided. The algorithm was trained over a GPU for about 50 iterations.The models training accuracy and validation accuracy increased during the duration of the training procedure with a total sample size of 64 and 50 epochs. The best validation accuracy of 98.57 was attained, and the best model was preserved by utilizing a checkpoint.In order to address slow convergence, the learning rate can be adjusted. The data was separated into sets for training and testing using the validation split of 0.2.The proportion ofsegments included in the validation sets and those used for the training were both identical in every class.In fig 18 and fig 19, we have also shown the plots for accuracy vs. epochs and loss vs. epochs. respectively.

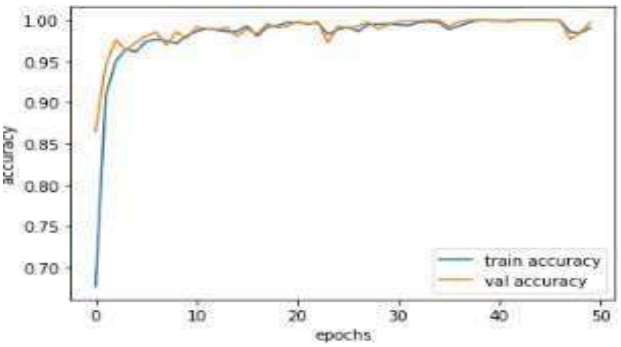


Fig. 18. graph for accuracy vs epochs

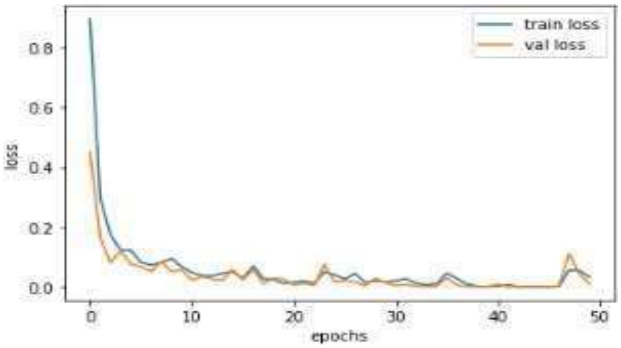


Fig. 19. graph for loss vs epochs

## V. CONCLUSION

Over the course of our work, we conducted research and read a range of study articles. The experiment was carried out to see whether it would be possible to determine a person's psychological emotion reaction by listening to brief recordings of their speech. Our approach has a 99.73 percent accuracy rate. We chose deep learning for our project because it produces more accurate outcomes than machine learning.In this work, we identified various speech emotion sounds with deep learning algorithms on the emotional speech recognition dataset, which was

supplemented by exploratory data analysis to provide more insights into the classification process. We chose the LSTM model because it enables the system to efficiently process continuous input streams without raising the needed bandwidths. In order to optimally utilise the model parameters, the proposed model alters the usual design of the LSTM network.Despite we obtained improved results with our model, much more work in the domain of voice emotion identification need to be done. Further work may be done to improve the process and make advantage of larger datasets.This is done to make sure the algorithm can produce satisfying results across a variety of data sources and much more than the courses we have taken, with extremely high accuracy on validation sets, enhanced prediction confidence, and more reliance on real-world data. Even if the final accuracy is rather excellent, we will continue to investigate ways to make the process even better.

## REFERENCES

[1] PavolHarar1 ,RadimBurget, and Malay Kishore Dutta, "Survey on Speech Emotion Recognition with Deep Learning", 2017.

[2] Sandeep Kumar Pandey, H.S.Shekhawat, and S.R.M.Prasanna, "Survey on Deep Learning Techniques for Speech Emotion Recognition".

[3] RaoudhaYahiaCherif, Abdelouahab Moussaoui, Nabila Frahta, and Mohamed Berrimi,"Survey on Effective speech emotion recognition using deep learning approaches for Algerian dialect".

[4] M. El Ayadi, M.S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, vol. 44, no. 3, pp.572-587, 2011.

[5] TanviPuri, MukeshSoni, Gaurav Dhiman, Osama Ibrahim Khalaf, Malik alazzam, and IhtiramRazaKhan, "Survey on Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network".

[6] P. Shen, Z. Changjun, and X. Chen, "Automatic speech emotion recog- nition using support vector machine," Proceedings of 2011 International Conference on Electronic Mechanical Engineering and Information Technology, vol. 2, pp. 621–625.

[7] A. Khalil, W. Al Khatib, E. El-Alfy, and L. Cheded, "Anger detection in arabic speech dialogs," 2018 International Conference on Computing Sciences and Engineering (ICCSE), pp. 1–6, Mar. 2018.

[8] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machine," International Journal of Smart Home, vol. 6, no. 2, pp.101-108, 2012.

[9] S. Lalitha, A.Madhavan, B.Bhushan, and S. Saketh, "Speech emotion recognition," In Advances in Electronics, Computers and Communications (ICAECC), 2014 International Conference, IEEE pp. 1- 4, 2014, October.

[10] Y. Zhou, Y. Sun, J. Zhang, and Y. Yan, "Speech emotion recognition using both spectral and prosodic features," In 2009 International Conference on Information Engineering and Computer Science, IEEE, pp. 1-4, 2009, December.

[11] Moshika, A., Thirumaran, M., Natarajan, B., Andal, K., Sambasivam, G., &Manoharan, R. (2021).Vulnerability assessment in heterogeneous web environment using probabilistic arithmetic automata. IEEE Access, 9, 74659-74673.

[12] A. Al-Faham, and N. Ghneim, "Towards enhanced arabic speech emotion recognition: comparison between three methodologies," Asian J. Sci. Technol, vol. 7, no. 3, pp. 2665–2669, 2016.

[13] H. Dahmani, H. Hussein, B. Meyer-Sickendiek, and O. Jokisch, "Nat ural Arabic Language Resources for Emotion Recognition in Algerian Dialect," International Conference on Arabic Language Processing, pp. 18–33, Oct. 2019.

[14] J. Niu, Y.Qian, and K. Yu, "Acoustic emotion recognition using deep neural network," In Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on, IEEE, pp. 128- 132, 2014, September.

[15] H. M. Fayek, M. Lech, and L. Cavedon. "Towards real-time speech emotion recognition using deep neural networks," Signal Processing and Communication Systems (ICSPCS), 2015 9th International Conference on. IEEE, 2015.

[16] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," In 22nd International Conference on Data Engineering Workshops (ICDEW'06), IEEE, pp. 8-8, 2006, April.

[17] P. Jackson, and S. Haq, "Surrey Audio-Visual Expressed Emotion(SAVEE) Database".

[18] J. Kim, and R. Saurous, "Emotion Recognition from Human Speech Using Temporal Information and Deep Learning," Interspeech, pp. 937– 940, Sep. 2018.

[19] Y. Hifny, and A. Ali, "Efficient arabic emotion recognition using deep neural networks," ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6710–6714, May 2019.

[20] A. Meftah, Y. Alotaibi, and S. Selouani, "Designing, building, and analyzing an arabic speech emotional corpus: Phase 2," 5th International Conference on Arabic Language Processing, pp. 181–184, Nov. 2014.

[21] HadhamiAouani and YassineBenayed, "Survey on Speech Emotion Recognition with deep learning," Multimedia InfoRmation systems and Advanced Computing Laboratory, MIRACL University of Sfax, Tunisia, 2 October 2020.

[22] F. Burkhardt, A.Paeschke, M.Rolfes, W.F.Sendlmeier, and B. Weiss, "A database of German emotional speech," In Interspeech, vol. 5, pp. 1517-1520, 2005, September.

[23] Rajesh, M., &Sitharthan, R. (2022). Image fusion and enhancement based on energy of the pixel using Deep Convolutional Neural Network. Multimedia Tools and Applications, 81(1), 873-885.

[24] Google WebRTC. https://webrtc.org/. Accessed 10 Oct 2016

[25] G.E. Hinton, N. Srivastava, A. Krizhevsky, I.Sutskever, and R.R. Salakhutdinov, "Improving neural networks by preventing coadaptation of feature detectors,"arXiv preprint arXiv:1207.0580, 2012.