IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 9, Issue 11, November 2020

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 11, November 2020||

Speech Emotion Recognition with MLP using Modified RAVDESS Dataset

Vrushali Dalvi[1]

IT Student, Department of Information Technology, B.K. Birla College of Arts, Science and Commerce (Autonomous), Kalyan, Maharashtra, India[1]

ABSTRACT: The field of machine learning has found its way in many real-world applications including Speech Emotion Recognition. Human beings convey their feelings through different forms including speech, expressions, gestures, and sometimes even silence. Out of all these forms speech is considered as a relevant and quicker form of expressing feelings or emotions of an individual. The field of recognizing emotions through speech has witnessed lot of research work by researchers around the world, yet it was found that it is a difficult task to identify the best approach towards developing such a system. After extensive literature survey, it was found that the most used classifiers were SVM and CNN. In this paper, a speech emotion recognition approach towards RAVDESS dataset with a slight modification is carried out using MLP to study its effect shifting from the most used SVM or CNN classifiers. This work basically tries to experiment the impact of MLP on a small dataset with 1440 audio samples and then modifies the dataset size for acquiring greater accuracy. The emotions under consideration for this work are happy, calm, sad, angry, surprised, fearful, disgust, and neutral.

KEYWORDS: Speech Emotion Recognition, feature extraction, feature classification, Human-Computer Interaction, Multi-Layer Perceptron

I. INTRODUCTION

SER has gained a lot of importance in the research field in this decade. The Speech Emotion Recognition is proven to be a very useful field when it comes to HCI. It can be used to know the emotions of a driver and take safety measures accordingly. Call centers also use speech emotion recognition to know the emotions of the customers and many more. Several techniques involved in the process are signal preprocessing, feature extraction, feature classification, etc. Humans find speech as the most effective way to express themselves and we realize this while trying to convey our thoughts through texts, emails, emojis, etc. Emotions help us to understand each other in a much better way so this fact is implemented in computer interaction as well. HCI helps computers to interact with humans and understand them and take actions accordingly. There are various research and studies in order to rectify the problems underlying in gaining higher accuracy in SER models as it is still an open problem, yet the discrete most common methods are mostly used. Various aspects and variables affect the recognition process such as gender and age as specified in Verma et.al. [6][16]. Most of the recent studies in SER seem to have focused on SVM for classification of emotions [1][3][5][6][8][14][16][20][21][23][24][26][27] [28][29][30[31]. After SVM, the second-best performing classifier was Decision Trees [8][27]. CNN was also used extensively in many studies involving SER [7][28].

The goal of this paper is to study the aspects of the MLP classifier for a change by shifting our focus from SVM, CNN, etc. and find out its working on a small dataset to make proper analysis about the relation between the MLP classifier and a small-sized dataset. MFCCs features were used in the proposed study which was extracted through feature extraction. The basic emotions included in this study are happiness, calm, anger, sadness, fear, surprise, disgust and neutral which are all the emotions in the dataset i.e. RAVDESS. This dataset consists of audio samples in English language.

The remaining paper is organized as follows: section II covers the main objectives of this study, section III consists of related SER studies, sections IV, V include the methodology and experimental results of the proposed study respectively, and finally section VI concludes the study along with future scope.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 11, November 2020||

II. RELATED WORK

Previous related works under the SER field are as follows: In [1] Peerzade et. al. contributed basic knowledge of speech emotion recognition system and different feature extraction and classification techniques for the system. Features were classified as Elicited features, Prosodic features, and Spectral features. Different classifying techniques were used to classify different emotions from human speech like Hidden Markov Model (HMM), Gaussian Mixtures Model (GMM), Support Vector Machine (SVM), Artificial Neural Network (ANN), K-nearest neighbor (KNN). Sepedi (A south African language) database was used to test the various speech recognition algorithms by Manamela et. Al. in [5].The three standard ML algorithms SVM, KNN and MLP were used. Instead Auto-WEKA algorithm was also used to test the accuracies. Neumann et.al. in [7] in their proposed work used representation learning and a labeled dataset was used for training the model for speech emotion recognition. Tests were performed on an unlabeled dataset and it was seen that using cross-corpus testing on CNN (Convolutional Neural Network) accuracy of the model increased by a considerable percentage. In [14] five different algorithms were used, they are: Logistic Regression (LR), K-Nearest Neighbour (KNN), Naive Bayesian classifier (B), Support Vector Machine (SVM) and, Multilayer Perceptron (MLP) of Neural Network. The two techniques that were used for feature extraction were Mel-scaled power spectrum and Mel frequency cepstrum coefficients. The dataset chosen was Surrey Audio-Visual Expressed Emotion Database i.e. SAVEE. Amongst the five algorithms used, Naive Bayesian classifier (B) showed the best results i.e. the highest accuracy compared to the remaining four algorithms. The SVM (Support Vector Machine), CNN (Convolutional Neutral Network) and LSTM-CNN (Long Short-Term Memory-Recurrent Neutral Network) were the three classifiers used in this study. SVM algorithm was used to prove the hypothesis whereas, in [16] , the SER system was distinguished on basis of gender recognition first. After signal preprocessing and acoustic feature extraction, the features were then classified based on gender which helped in improving the accuracy. The average accuracy of the proposed work was 78% using SVM. The other two algorithms that were used were KNN and MLP.

III. METHODOLOGY

1) Datasets included:

- RAVDESS:

In our proposed study we use The Ryerson Audio-Visual Database of Emotional Speech and Song. It was collected from ‘Kaggle’. [32] This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically matched statements in a neutral North American accent. Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

2) Proposed model

By reviewingrelated studies, it can be concluded that every classification algorithm has its own advantages and disadvantages. SVM algorithm is mostly used in recognition of emotions through speech because it is faster than other algorithms but, SVM requires more and more data to perform better.Hence, this study shifts its focus to another classification algorithm i.e. MLP (Multi-Layer Perception). Firstly, the feature extraction process is carried out to extract the MFCCs from the audio samples. 40 MFCCs were returned by the feature extraction. The proposed model was trained with a learning rate (alpha) of 0.01 with a batch size of 256. The number of hidden layers was set to 300. The maximum iterations were set to 500. First, we train the model using audio files from 'Actor_01-Actor_16'. The total number of files in these folders were 960. The dataset was then split into a train test split of 8:2 ratio. This size of the RAVDESS resulted into less accuracy which is discussed in the experimental results section. The hyper-parameters of the proposed model were adjusted to check if it affects the accuracy. The maximum increase noted was an increase of about 7%. So, the original dataset size was modified. The original dataset consisted of 1440 audio files (.wav format). The dataset size was doubled by adding 1440 additional audio speech files from another source [33].The hyper-parameter values were kept the same as mentioned above but this time we had double the data. Figures 3 and 4 indicate waveplots of speech signals for the emotions fear and happiness. The difference in the amplitude and frequencies of these speech signals can be clearly seen with the help of the speech signal. Figures 5(a)-(d) demonstrate the angry emotion speech signals for female and male

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 11, November 2020||

gender respectively along with their MFCC colorbar plots. Similarly figures 6 (a)-(d) show the speech signal waveplots and MFCC colorbar plots for the happy emotion. All these plots are plotted with respect to time. These two emotions are opposites of one another, so the distinguishing factor is also more.

•MFCC:

MFCCs are derivatives of a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression. Figure(2) represents a block diagram of the processes involved in MFFC extraction. The librosa and libsvm were used for implementing our proposed model.

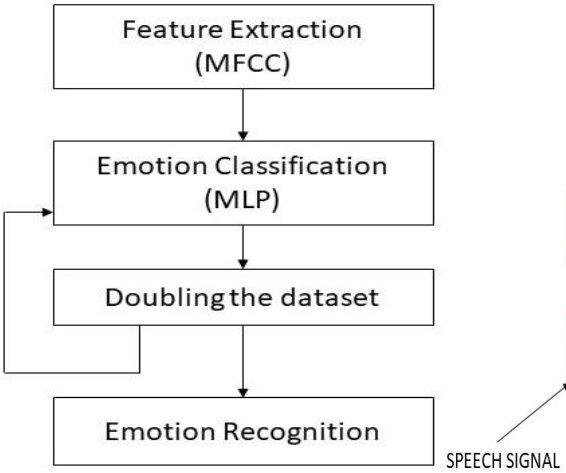


Fig. 1: Block Diagram of Proposed Model

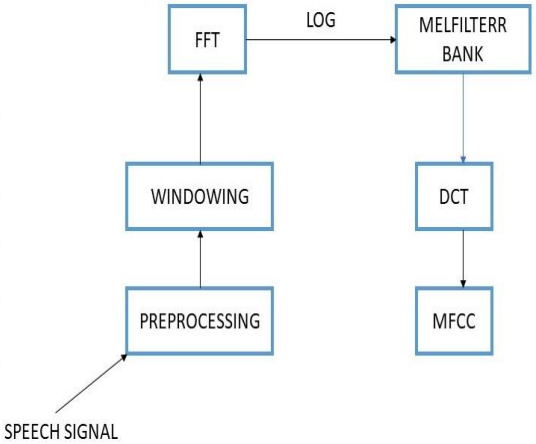


Fig. 2: MFCC extraction

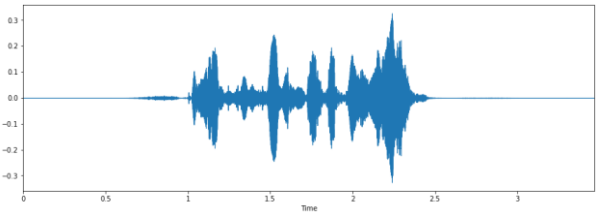


Fig. 3: Waveplot for speech signal(fear)

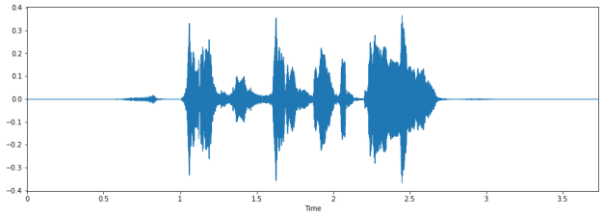
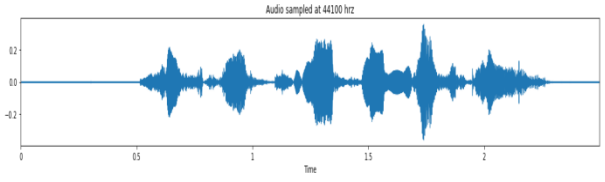
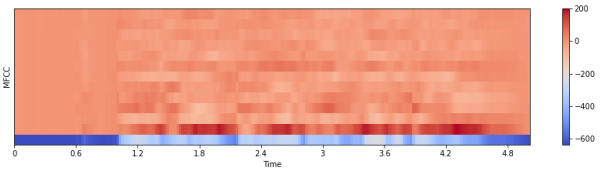


Fig. 4 Waveplot for speech signal(happy)



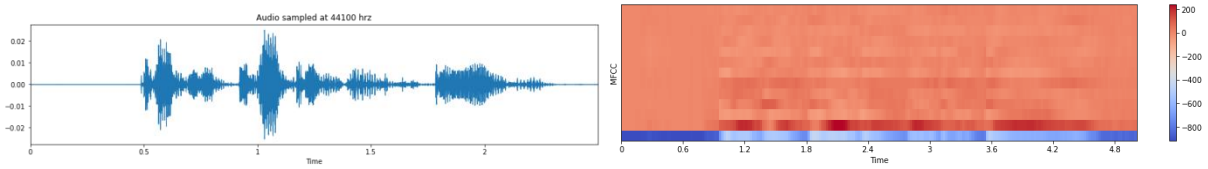
(a)



(b)

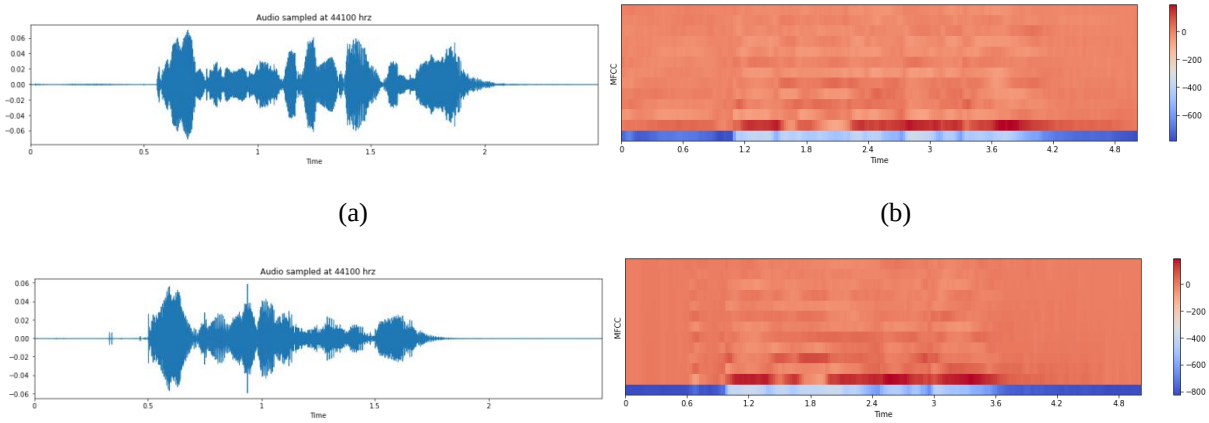
| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 11, November 2020||



(c) (d)

Fig. 5 a) Waveplot for speech signal(angry-female), (b) MFCC colorbar plot(angry-female), (c) Waveplot for speech signal(angry-male), (d) MFCC colorbarplot(angry male)



(a) (b)

(c) (d)

Fig. 6(a) Waveplot for speech signal(happy-female), (b) MFCC colorbar plot(happy-female), (c) Waveplot for speech signal(happy-male), (d) MFCC colorbar plot(happy-male)

IV. EXPERIMENTAL RESULTS

Our study proposed a model where we first perform our experiments on a small dataset with 1440 speech samples. These were split into a train-test split of 80% and 20%. The hyper-parameter values during this predication step were as follows: learning rate (alpha=0.01), batch size was set to 256, number of hidden layers was set to 300, and the maximum iterations were set to 500. The accuracy obtained on the test set 33.33% with this size of the dataset. Further we adjusted the hyper-parameter values by changing the batch size to 200, and number of hidden layers to 600 i.e. twice the previous number. The accuracy at this step increased by a value of about 7%. The accuracy obtained by performing these adjustments was 40.89%. But this is not a satisfactory percentage. So, the dataset size was doubled keeping the hyper-parameters same. By doing this, the proposed model achieved an accuracy of 88.04% which is a reasonably good accuracy. Our proposed model gave an average precision-recall score of 0.90. The roc-auc score for our proposed model was 0.957. Table 1 indicates a simple summary of our experiments. Table 2 summarizes a comparison between the previous related model accuracies and our proposed model's accuracy. Figure 7 demonstrates a bar graph indicating a rise and fall in the accuracies with respect to the dataset size based on our experiments.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 11, November 2020||

Study	Reference No.	Accuracy (%)
The Automatic Recognition of Sepedi Speech Emotions Based on Machine Learning Algorithms	[5]	59.9, 48.8, 51.8
Performance Analysis of ML Algorithms on Speech Emotion Recognition	[14]	83.3
Role of gender influence in vocal Hindi conversations: A study on speech emotion recognition	[16]	68.5
Proposed model	-	88.04

Table. 1 Comparison of proposed model accuracy with previous related studies.

Model	Size of Dataset	Hyperparameter values	Accuracy
1	1440	alpha=0.01, batch size=256, no. of hidden layers=300, max_iter=500	33.33%
2	1440	alpha=0.01, batchsize=200, no.ofhidden layers=600, max_iter=500	40.89%
3	2880	alpha=0.01, batchsize=200, no.ofhidden layers=600, max_iter=500	88.04%

Table. 2 Summary of experiments and accuracy results

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 11, November 2020||

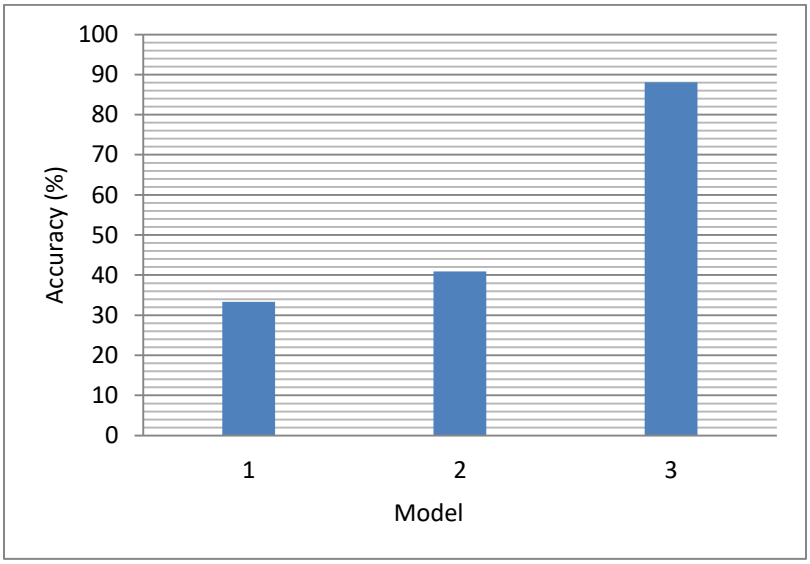


Fig. 7 Trends in accuracies as per our experiments (based on table 2)

V. CONCLUSION

In our study, we proposed a model a model using the MLP classifier. MFCCs were extracted through feature extraction. We experimented with modifying the size of the dataset to check the changes in our accuracy scores. It can be concluded that like SVM, MLP classifier also need more data but the size requirement is still less compared to the SVM classifiers. Doubling the dataset size also doubled the accuracy rate. So the dataset size is directly proportional to the accuracy rate and could indicate a straight line when plotted on a graph. In future, this model can be improved by making the dataset times three times the original size to achieve a greater accuracy.

VI. ACKNOWLEDGEMENT

I would like to thank Prof. Swapna Augustine Nikale, Department of Information Technology, B.K Birla College Kalyan for providing guidance for this research work.

VII. GLOSSARY

SER- Speech Emotion Recognition

HCI- Human Computer Interaction

MFCC- Mel Frequency Cepstral Coefficients

RAVDESS- The Ryerson Audio-Visual Database of Emotional Speech and Song

CNN- Convolutional Neural Networks

MLP- Multi Layer Perceptron

SVM- Support Vector Machines

REFERENCES

[1] Peerzade, Gulnaz Nasir & Deshmukh, Ratnadeep& Waghmare, S.D.. (2018). A Review Speech Emotion Recognition. International Journal of Computer Sciences and Engineering. 6. 400-402. 10.26438/ijcse/v6i3.400402.

[2] Sekkate, S., Khalil, M., Adib, A., & Ben Jebara, S. (2019). An Investigation of a Feature-Level Fusion for Noisy Speech Emotion Recognition. Computers, 8(4), 91. https://doi.org/10.3390/computers8040091

[3] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Ali Mahjoub, M., &Cleder, C. (2020). Automatic Speech Emotion Recognition Using Machine Learning. Social Media and Machine Learning, 1–16. https://doi.org/10.5772/intechopen.84856

[4] Audiovisual Information Processing in Emotion Recognition: An Eye Tracking Study. (2020). Cognitivesciencesociety, 2625–2630. https://cognitivesciencesociety.org/cogsci20/papers/0634/0634.pdf

[5] Manamela, Phuti&Manamela, Madimetja&Modipa, Thipe&Sefara, Tshephisho&Mokgonyane, Tumisho. (2018). The Automatic Recognition of Sepedi Speech Emotions Based on Machine Learning Algorithms. 10.1109/ICABCD.2018.8465403.

[6] Verma, D., & Mukhopadhyay, D. (2016a). Age driven automatic speech emotion recognition system. 2016 International Conference on Computing, Communication and Automation (ICCCA), 1005–1010. https://doi.org/10.1109/ccaa.2016.7813862

[7] Neumann, M., & Vu, N. T. (2019). Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7390–7394.

[8] Koduru, A., Valiveti, H.B. &Budati, A.K. Feature extraction algorithms to improve the speech emotion recognition rate. Int J Speech Technol 23, 45–55 (2020). https://doi.org/10.1007/s10772-020-09672-4

[9] Wang, C. (2020, January 16). Speech Emotion Recognition Based on Multi-feature and Multi-lingual Fusion. ArXiv.Org. https://arxiv.org/abs/2001.05908

[10] Tarantino, L., Garner, P. N., & Lazaridis, A. (2019). Self-Attention for Speech Emotion Recognition. Interspeech 2019, 2578–2582. https://doi.org/10.21437/interspeech.2019-2822

[11] A Hybrid approach for speech emotion recognition using 1D-CNN LSTM. (2020). Researchgate, 833–835. https://www.researchgate.net/profile/Tosin_Adesuyi/publication/342765156_A_Hybrid_approach_for_speech_emotion_recognition_using_1D-CNN_LSTM/links/5f058611a6fdcc4ca455d8f7/A-Hybrid-approach-for-speech-emotion-recognition-using-1D-CNN-LSTM.pdf

[12] Subhashree, R., &Rathna, G. N. (2016). Speech Emotion Recognition: Performance Analysis based on Fused Algorithms and GMM Modelling. Indian Journal of Science and Technology, 9(11), 1–8. https://doi.org/10.17485/ijst/2016/v9i11/88460

[13] Latif, S. (2018, January 19). Transfer Learning for Improving Speech Emotion Classification Accuracy. ArXiv.Org. https://arxiv.org/abs/1801.06353

[14] Tiwari, P., &Darji, A. D. (2020). Performance Analysis of ML Algorithms on Speech Emotion Recognition. Advances in Intelligent Systems and Computing, 91–100. https://doi.org/10.1007/978-981-15-1275-9_8

[15] Chakraborty, R., Pandharipande, M., &Kopparapu, S. K. (2016). Knowledge-based Framework for Intelligent Emotion Recognition in Spontaneous Speech. Procedia Computer Science, 96, 587–596. https://doi.org/10.1016/j.procs.2016.08.239

[16] Verma, D., Mukhopadhyay, D., & Mark, E. (2016). Role of gender influence in vocal Hindi conversations: A study on speech emotion recognition. 2016 International Conference on Computing Communication Control and Automation (ICCUBEA), 1–7. https://doi.org/10.1109/iccubea.2016.7860021

[17] Sahu, S. (2018, June 18). On Enhancing Speech Emotion Recognition using Generative Adversarial Networks. ArXiv.Org. https://arxiv.org/abs/1806.06626v1

[18] Zhang, Y., Du, J., Wang, Z., Zhang, J., & Tu, Y. (2018). Attention Based Fully Convolutional Network for Speech Emotion Recognition. 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 1–5. https://doi.org/10.23919/apsipa.2018.8659587

[19] Peng, Z., Li, X., Zhu, Z., Unoki, M., Dang, J., & Akagi, M. (2020). Speech Emotion Recognition Using 3D Convolutions and Attention-Based Sliding Recurrent Networks With Auditory Front-Ends. IEEE Access, 8, 16560–16572. https://doi.org/10.1109/access.2020.2967791

[20] Elbarougy, R. (2019). Speech Emotion Recognition based on Voiced Emotion Unit. International Journal of Computer Applications, 178(47), 22–28. https://doi.org/10.5120/ijca2019919355

[21] Pakyurek, M., Atmis, M., Kulac, S., &Uludag, U. (2020). Extraction of Novel Features Based on Histograms of MFCCs Used in Emotion Classification from Generated Original Speech Dataset. ElektronikaIrElektrotechnika, 26(1), 46–51. https://doi.org/10.5755/j01.eie.26.1.25309

[22] Langari, S., Marvi, H., & Zahedi, M. (2020). Efficient speech emotion recognition using modified feature extraction. Informatics in Medicine Unlocked, 20, 100424. https://doi.org/10.1016/j.imu.2020.100424

[23] Jain, M. (2020, February 3). Speech Emotion Recognition using Support Vector Machine. ArXiv.Org. https://arxiv.org/abs/2002.07590v1

[24] Pa Pa Win, H., & Thu ThuKhine, P. (2020). Emotion Recognition System of Noisy Speech in Real World Environment. International Journal of Image, Graphics and Signal Processing, 12(2), 1–8. https://doi.org/10.5815/ijigsp.2020.02.01

[25] Impact of Dithering in Compressed and Spectrally Distorted Speech for Emotion Recognition. (2020). International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST), 25–28. https://doi.org/10.20238/IJARBEST.2020.0610005

[26] Sharma, U. (2020). Detection of Depression from Speech Signal through Linear SVM using MFCC Feature. SSRN Electronic Journal, 1–4. https://doi.org/10.2139/ssrn.3562977

[27] R. (2017). RahulYerramsetti/Speech-Emotion-Recognition-Using-SVM-Decision-Tree-and-LDA. GitHub. https://github.com/RahulYerramsetti/Speech-Emotion-Recognition-Using-SVM-Decision-Tree-and-LDA

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 11, November 2020||

[28] I. (2018). ireneCccc/Speech-Emotion-Recognition. GitHub. https://github.com/ireneCccc/Speech-Emotion-Recognition

[29] Quan, C., Zhang, B., Sun, X., & Ren, F. (2017). A combined cepstral distance method for emotional speech recognition. International Journal of Advanced Robotic Systems, 14(4), 172988141771983. https://doi.org/10.1177/1729881417719836

[30] Liu, Z.-T., Wu, M., Cao, W.-H., Mao, J.-W., Xu, J.-P., & Tan, G.-Z. (2018). Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing, 273, 271–280. https://doi.org/10.1016/j.neucom.2017.07.050

[31] Zhu, L., Chen, L., Zhao, D., Zhou, J., & Zhang, W. (2017). Emotion Recognition from Chinese Speech for Smart Affective Services Using a Combination of SVM and DBN. Sensors, 17(7), 1694. https://doi.org/10.3390/s17071694

[32] https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio

[33] https://zenodo.org/record/1188976#.X6nlp3QzbIU

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details