IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Speech Emotion Recognition

Mr. M China Pentu Saheb[1], Assistant Professor, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.

P Sai Srujana[2], **P Lalitha Rani**[3], **M Siva Jyothi**[4]
[2,3,4] UG Students, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
[1]saheb10thjune@gmail.com, [2]psaisrujana2001@gmail.com,
[3]Lalitharani.palakaluri@gmail.com, [4]jyothimeda199@gmail.com

DOI:10.48047/IJFANS/V11/I12/203

Abstract
Emotions are the best way for people to communicate their thoughts and actions to others. The most important technology in the world today is the ability to recognize emotions from a single speaker's voice. The ability to recognize emotions is very useful in gaining various insightful insights into a person's thoughts. The process of extracting emotions from human speech is called Speech Emotion Recognition (SER). We used the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset to extract emotions from Speech. Emotions are extracted from speech based on speech parameters such as Mel-Frequency-Cepstral -Coefficients (MFCC) and Mel Spectrogram. After training with a Multilayer Perceptron classifier (MLP), the obtained data had an accuracy of 68.33% and accuracy of 80.64% after training with Convolutional Neural Networks Long Short Term Memory (CNN LSTM).

Keywords: CNN LSTM, Mel, MFCC, MLP, SER

Introduction

Speech emotion recognition is one of the fastest growing research topics in the world of computer science. Emotions are the mediums used to describe how a person is feeling and their state of mind. Emotions play an important role in sensitive professions such as surgeons, military commanders, and many other professions that require you to keep your emotions in check. There are multiple ways the expresses human emotions, including posture, facial expressions, and voice. Out of these, speech very important for emotional expression. To communicate effectively with people, the system needs to understand the emotion of speech. Many machine learning algorithms are developed and tested to classify the emotions. The vital part of speech emotion recognition is Feature extraction. Feature equality directly affects the accuracy of the classification results. Predicting emotions is a difficult task. This work is carried out to recognize the emotions in the given audio samples. We plan to use MLP and CNN LSTM to recognize emotions. We compare the two

classifiers. CNN LSTM consistently and efficiently predicts the emotion of speech input compared to MLP.

Literature Survey

[1]Peipei Shen et al provided an automatic Speech Emotion recognition using support Vector machine and accomplished an accuracy of 66.02%. [2]Seyedmahdad Mirsamadi et al proposed automatic Speech Emotion recognition using Recurrent Neural Networks with nearby interest and acquired an accuracy of 63.5%. [3]Puneet Kumar et al presented multimodal Speech Emotion recognition to identify 4 feelings from speech using Gated Recurrent devices(GRU) and Bidirectional Encoder Representations from Transformers(BERT) and achieved an accuracy of 71%. [4]Chi-Chun Lee et al proposed Emotion recognition the usage of a Hierarchical Binary decision Tree approach to recognize four feelings from speech which obtains development over the SVM baseline.[5]Mao Li et al presented Contrastive Unsupervised learning for Speech Emotion recognition to recognize only two emotions like anger and sadness using Contrastive Predictive Coding (CPC).

Problem Identification

Recognizing emotions from speech is a rapidly growing field of research that aims to develop automatic systems capable of detecting and interpreting emotional states from speech signals. While recent advances in machine learning led to significant improvements in SER performance, there are still several challenges that need to be addressed to make these systems more accurate and efficient.

Some of the specific challenges associated with speech emotion recognition include dealing with the changes in emotional expression across different speakers, accounting for the influence of background noise on speech signals, and addressing issues related to data quality .

Furthermore, there is also a need to develop robust and interpretable models that can generalize well across different contexts and be easily integrated into practical applications, such as automated customer service systems, mental health support platforms, and virtual assistants. This project is aimed at developing a system that can accurately recognize the emotional state of a speaker based on their speech signal. [21-29]

Methodology

The models MLP and CNN LSTM are used to recognize emotions from speech using the RAVDESS dataset. RAVDESS dataset consists of 2800 speech files collected from 24 actors numbered from 01 to 24.Out of 24 actors there are 12 male actors represented with odd numbers and 12 Female actors represented with even numbers. Our model is trained using

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

audio-only data because our main goal is to recognize emotions in speech. All 24 actors vocalize two predetermined statements for each of the eight emotions repeated twice for each sentence except the "neutral" emotion, which has only normal intensity. There are two levels of intensity for all emotions i.e., normal and strong. The labels for the emotions are represented with the numbers 01 to 08.

The Key features of the audio data are extracted using MFCC (Mel Frequency Cepstral Coefficients) and Mel Spectrogram. MFCC is an important feature extraction technique used when using audio datasets. Mel scale is a scale that relates the perceived frequency of a tone to the real measured frequency. It scales the frequency so that you can fit greater carefully what the human ear can hear. Mel Spectrogram is obtained by applying a fast Fourier transform on overlapping segments of the signal, and spectrogram is visual way of representation of signal strength and also used to display the frequency of sound waves. A Mel Spectrogram is displayed in Fig 1.

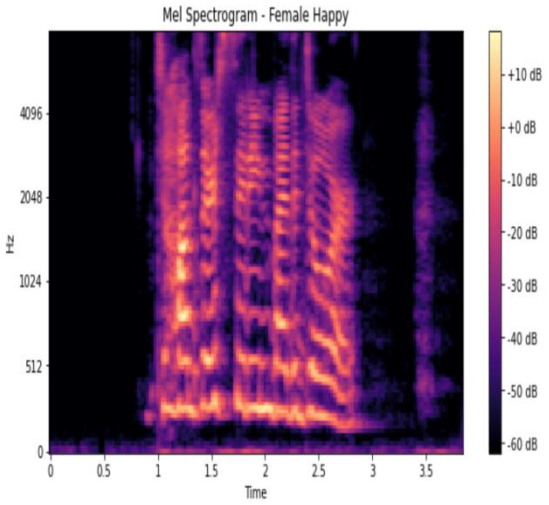

Fig1: Mel Spectrogram for female_happy

A Multi Layer Perceptron consists of three types of layers. They are input layer, hidden layer and output layer. The initial data is received by the input layer and then multiplied with weights to create the hidden layer. Nonlinear data is learned by the model through activation functions in this layer. The hidden layers continue to process the calculation until the final output layer. The activation function utilized for our application is the rectified linear unit activation function. The hidden layer is not directly exposed to the input. The output layer produces a value or vector in the required format for the problem.

CNN LSTM, is an LSTM structure that is tailored to solving sequence prediction challenges. To create a CNN LSTM, CNN layers are incorporated at the beginning, then LSTM layers are added, and finished with a dense layer on the output. The model can be thought of as

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

having two distinct components: the CNN Model, which performs feature extraction, and the LSTM Model, which analyzes the extracted features. The CNN layers extract features from the input data, while LSTMs handle sequence prediction.

Implementation

To implement Speech Emotion Recognition using MLP and CNN LSTM the default split ratio is used to split the dataset into training and testing datasets i.e., 70% for training dataset and 30% for testing dataset. For Exploratory Data Analysis MFCC and Mel Spectrogram are used. The count of emotions for each category is shown in Fig2.

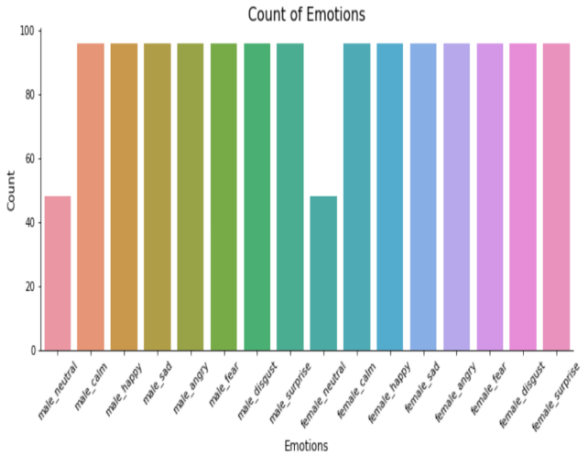

Fig 2: Count of Emotions

Data augmentation is performed by adding small disturbance on our initial training dataset. Polymerized data for audio, is generated by applying different data augmentation techniques. The goal is to make our model insensitive to such disruption and improve its generalizability. To make this work, the disturbance must have the same label as the original training sample. The features are cepstral coefficients from zeroth order coefficient to thirteenth order coefficient followed by delta of the coefficients until delta of fifth order coefficient which are 20 in number and are extracted into into a csv file.

MLP is implemented with a hidden layer consisting of 2300 neurons and obtained an accuracy of 68.33% on the testing dataset. CNN LSTM is implemented by using three one dimensional convolutional layers each accompanied with a batch normalization layer and a max pooling layer. The activation function used for all the convolutional layers is Rectified Linear Unit. There are two LSTM layers followed by three dense layers. Finally there is a fully connected layer at the end. Softmax function is also used.

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Results & Conclusion

The emotions from speech are classified into 16 classes out of which 8 classes are female emotions and 8 classes are male emotions as listed in Fig 4.

MLP achieved an accuracy of 68.33% on the training dataset and CNN LSTM achieved an accuracy of 80.64%. The confusion matrix and performance metrics of CNN LSTM are represented in Fig3 and Fig4.

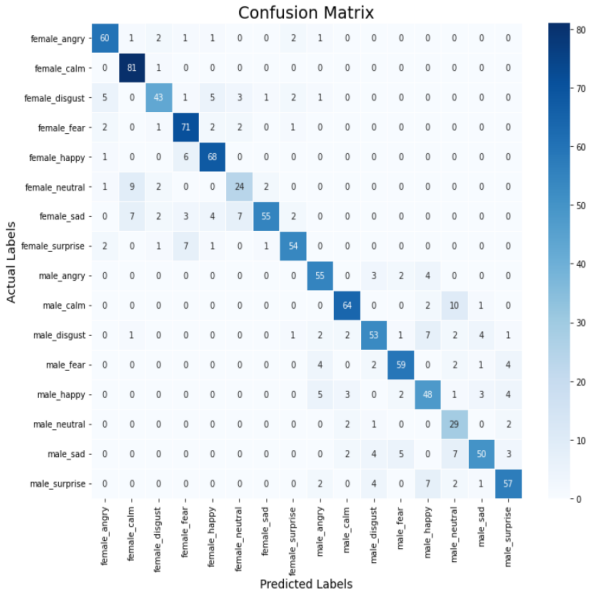

Fig3: Confusion Matrix for CNN LSTM

	precision	recall	f1-score	support
female_angry	0.85	0.88	0.86	68
female_calm	0.82	0.99	0.90	82
female_disgust	0.83	0.70	0.76	61
female_fear	0.80	0.90	0.85	79
female_happy	0.84	0.91	0.87	75
female_neutral	0.67	0.63	0.65	38
female_sad	0.93	0.69	0.79	80
female_surprise	0.87	0.82	0.84	66
male_angry	0.79	0.86	0.82	64
male_calm	0.88	0.83	0.85	77
male_disgust	0.79	0.72	0.75	74
male_fear	0.86	0.82	0.84	72
male_happy	0.71	0.73	0.72	66
male_neutral	0.55	0.85	0.67	34
male_sad	0.83	0.70	0.76	71
male_surprise	0.80	0.78	0.79	73
accuracy			0.81	1080
macro avg	0.80	0.80	0.80	1080
weighted avg	0.81	0.81	0.81	1080

Fig4: Performance Metrics for CNN LSTM

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

The performance of Multi-Layer Perceptron (MLP) and Convolutional Neural Network Long Short-Term Memory (CNN LSTM) models are compared for speech emotion recognition. Our experiments showed that the CNN LSTM model works better than the MLP model, achieving a higher accuracy rate and demonstrating its superiority in processing sequential data.

The results of our study conclude that the CNN LSTM model is a promising approach for speech emotion recognition tasks.

Limitations & Future Scope

MLP and CNN LSTM models can detect emotion based on the acoustic features of speech, but do not consider contextual information, which is important for understanding emotion. For example, sarcasm can be difficult to detect with these models because they need to understand the context. Including facial expressions and gestures can provide additional cues for emotion detection. Combining information from voice, facial expressions, and physiological cues, can enhance the accuracy of emotion detection in real-world scenarios.

References

[1] Shen, P., Changjun, Z., & Chen, X. (2011, August). Automatic speech emotion recognition using support vector machine. In Proceedings of 2011 international conference on electronic & mechanical engineering and information technology (Vol. 2, pp. 621-625). IEEE.

[2] Mirsamadi, S., Barsoum, E., & Zhang, C. (2017, March). Automatic speech emotion recognition using recurrent neural networks with local attention. In 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP) (pp. 2227-2231). IEEE.

[3] Kumar, P., Kaushik, V., & Raman, B. (2021). Towards the Explainability of Multimodal Speech Emotion Recognition. In Interspeech (pp. 1748-1752).

[4] Lee, C. C., Mower, E., Busso, C., Lee, S., & Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. Speech Communication, 53(9-10), 1162-1171.

[5] Li, M., Yang, B., Levy, J., Stolcke, A., Rozgic, V., Matsoukas, S., ... & Wang, C. (2021, June). Contrastive unsupervised learning for speech emotion recognition. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6329-6333). IEEE.

[6] Lin, Y. P., Wang, C. H., Jung, T. P., Wu, T. L., Jeng, S. K., Duann, J. R., & Chen, J. H. (2010). EEG-based emotion recognition in music listening. IEEE Transactions on Biomedical Engineering, 57(7), 1798-1806.

[7] Schaaff, K., & Schultz, T. (2009, September). Towards an EEG-based emotion recognizer for humanoid robots. In RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication (pp. 792-796). IEEE.

[8] Pierre-Yves, O. (2003). The production and recognition of emotions in speech: features and algorithms. International Journal of Human-Computer Studies, 59(1-2), 157-183.

[9] Damodar, N., Vani, H. Y., & Anusuya, M. A. (2019). Voice emotion recognition using CNN and decision tree. Int. J. Innov. Technol. Exp. Eng, 8, 4245-4249.

[10] Zhao, J., Mao, X., & Chen, L. (2018). Learning deep features to recognise speech emotion using merged deep CNN. IET Signal Processing, 12(6), 713-721.

[11] Palo, H. K., Mohanty, M. N., & Chandra, M. (2015). Use of different features for emotion recognition using MLP network. In Computational Vision and Robotics: Proceedings of ICCVR 2014 (pp. 7-15). Springer India.

[12] Livingstone, S. R., & Russo, F. A. (2018). 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), e0196391.

[13] Boigne, J., Liyanage, B., & Östrem, T. (2020). Recognizing more emotions with less data using self-supervised transfer learning. arXiv preprint arXiv:2011.05585.

[14] Venkataramanan, K., & Rajamohan, H. R. (2019). Emotion recognition from speech. arXiv preprint arXiv:1912.10458.

[15] Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. Neural Networks, 92, 60-68.

[16] Riera, P., Ferrer, L., Gravano, A., & Gauder, L. (2019). No sample left behind: Towards a comprehensive evaluation of speech emotion recognition system. In Proc. Workshop on Speech, Music and Mind 2019.

[17] Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59, 101894.

[18] Zhao, Y., Yin, D., Luo, C., Zhao, Z., Tang, C., Zeng, W., & Zha, Z. J. (2021). General-Purpose Speech Representation Learning through a Self-Supervised Multi-Granularity Framework. arXiv preprint arXiv:2102.01930.

[19] Maithri, M., Raghavendra, U., Gudigar, A., Samanth, J., Barua, P. D., Murugappan, M., ... & Acharya, U. R. (2022). Automated emotion recognition: Current trends and future perspectives. Computer methods and programs in biomedicine, 106646.

[20] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. IEEE Access, 7, 117327-117345.

[21] Sri Hari Nallamala, et al., "A Literature Survey on Data Mining Approach to Effectively Handle Cancer Treatment", (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7, SI 7, Page No: 729 – 732, March 2018.

[22] Sri Hari Nallamala, et.al., "An Appraisal on Recurrent Pattern Analysis Algorithm from the Net Monitor Records", (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7, SI 7, Page No: 542 – 545, March 2018.

[23] Sri Hari Nallamala, et.al, "Qualitative Metrics on Breast Cancer Diagnosis with Neuro Fuzzy Inference Systems", International Journal of Advanced Trends in Computer Science and Engineering, (IJATCSE), ISSN (ONLINE): 2278 – 3091, Vol. 8 No. 2, Page No: 259 – 264, March / April 2019.

[24] Sri Hari Nallamala, et.al, "Breast Cancer Detection using Machine Learning Way", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue-2S3, Page No: 1402 – 1405, July 2019.

[25] Sri Hari Nallamala, et.al, "Pedagogy and Reduction of K-nn Algorithm for Filtering Samples in the Breast Cancer Treatment", International Journal of Scientific and Technology Research, (IJSTR), ISSN: 2277-8616, Vol. 8, Issue 11, Page No: 2168 – 2173, November 2019.

[26] Kolla Bhanu Prakash, Sri Hari Nallamala, et al., "Accurate Hand Gesture Recognition using CNN and RNN Approaches" International Journal of Advanced Trends in Computer Science and Engineering, 9(3), May – June 2020, 3216 – 3222.

[27] Sri Hari Nallamala, et al., "A Review on 'Applications, Early Successes & Challenges of Big Data in Modern Healthcare Management'", Vol.83, May - June 2020 ISSN: 0193-4120 Page No. 11117 – 11121.

[28] Nallamala, S.H., et al., "A Brief Analysis of Collaborative and Content Based Filtering Algorithms used in Recommender Systems", IOP Conference Series: Materials Science and Engineering, 2020, 981(2), 022008.

[29] Nallamala, S.H., Mishra, P., Koneru, S.V., "Breast cancer detection using machine learning approaches", International Journal of Recent Technology and Engineering, 2019, 7(5), pp. 478–481.