

Application Based System for Recognizing Speech Emotions

¹Dr. Rachna Somkunwar, ²Mr. Anil Kumar Gupta, ³Ishika Jain, ⁴Neha Shilvant, ⁵Rashi Pandey, ⁶Shreyas Deshpande

^{1,3,4,5,6}Associate Professor, Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri Pune, Maharashtra, India

²Senior Member IEEE CDAC, Pune, Maharashtra, India

Abstract - Businesses often underestimate the power of customer care or customer support to grow their revenues. Customers often receive generic responses that do not address their specific needs or emotions, leading to a lack of connection and dissatisfaction. Long wait times and slow response times can be frustrating for customers and lead to decreased satisfaction. Moreover poorly trained customer service agents can struggle to handle complex customer inquiries and provide adequate support, leading to dissatisfaction. They may not be able to accurately detect the emotional state of the customer, leading to an inappropriate response and further dissatisfaction. This can lead to shutting down of businesses. This paper proposes a system to provide a more personalized and empathetic response to customers by building a model using MLP Classifier. We are optimistic that our system based on MLP Classifier is more reliable as compared to the rest of the models available currently.

Keywords: MLP, Customer Satisfaction, Speech Signals.

I. INTRODUCTION

In the present day, customer satisfaction is a key concern for businesses and organizations. The traditional means of gauging it include conducting surveys and distributing questionnaires. Nevertheless, businesses and marketers are searching for faster and more efficient methods to gather feedback from their potential customers.[1] Customer support is a crucial aspect of any business and plays a vital role in maintaining customer satisfaction and loyalty. However, delivering effective customer support can be challenging, especially when dealing with a high volume of calls. A significant portion of these calls involve customers expressing their emotions, which can be difficult to accurately interpret, leading to misunderstandings and dissatisfied customers. This problem has become increasingly significant with the rise of online commerce and remote customer support services.

The goal is to accurately analyze the emotions of customers from their speech data. The problem is that traditional methods are vulnerable to interference from factors

like speaker differences and environmental noise, and the data has unbalanced sample classes.[2]

In this research paper, we propose a speech emotion recognition system based on Multilayer Perceptron (MLP), a type of artificial neural network. The MLP architecture is well-suited for speech emotion recognition because of its ability to learn complex non-linear relationships between the input features and the emotional state of the speaker. The proposed system is trained on a large dataset of speech samples, labelled with the corresponding emotions. The performance of the system is evaluated using standard evaluation metrics such as accuracy, precision, recall, and F1-score.

In detail, our system first pre-processes the speech signals to extract relevant acoustic and prosodic features. These features are then fed into the MLP, which consists of multiple layers of interconnected nodes, to determine the emotional state of the speaker. The MLP is trained using a supervised learning approach, where the network is trained on the pre-processed speech signals, labelled with the corresponding emotions. The parameters of the MLP are optimized using a back propagation algorithm to minimize the prediction error.

Once the MLP is trained, it can be used to classify speech signals into one of several emotional categories, such as happy, sad, angry, and neutral. The performance of the system is evaluated using a held-out test dataset, which contains speech signals that were not used during the training phase.

The evaluation results provide insight into the accuracy and robustness of the proposed system and highlight any areas that need further improvement.

II. LITERATURE REVIEW

In [1], authors have proposed a new method for facial emotion detection to recognize customer's satisfaction using machine-learning techniques with the used a facial landmark point; by extracting geometric features from customer's emotional faces using distances from landmarks points, they have used distances between the neutral side and the negative or positive feedback. After that, they classified these distances

by using different classifier, namely Support Vector Machine (SVM), KNN, Random Forest, Adaboost, and Decision Tree. To assess the method, verification of algorithm is done by using JAFFE datasets Advantage of their model is that it predicts the facial emotions fast and its accuracy is higher as compared to some of the pre-existing methods. Disadvantage is that if the amount of data is quite large hence algorithm takes a lot of time and may cause over fitting.

In [2], authors have proposed a method to calculate the spectrogram of speech and its first-order and second-order difference, stack the three as the input of neural network to reduce the influence of emotion independent factors; they have used CNN and LSTM to extract speech data features, and add attention mechanism to make the model focus on the time-frequency region related to emotion. For the problem of unbalanced sample classes, they have added Focal Loss to reduce the weight of samples which are easy to classify in loss. The results of experiment shows that the recognition accuracy of our model is 92.60% of weighted accuracy and 92.02% of unweighted accuracy, which is significantly improved compared with the traditional DNN-ELM method. Disadvantage is that if the amount of data is quite large then LSTM takes a lot of time and may cause over fitting on the other hand DNN does not contain memory unit it cannot handle sequence of data.

In [3], the authors investigated two distinct methods of feature extraction for effective speech emotion recognition. Initially, they proposed a two-way feature extraction technique that employs super convergence to extract two sets of potential features from the speech data. The first set of features was obtained using principal component analysis (PCA), which was then used as input to a deep neural network (DNN) containing dense and dropout layers. In the second approach, mel-spectrogram images were extracted from audio files and fed into a pre-trained VGG-16 model. The experimental results showed that the proposed models outperformed existing models in terms of various performance metrics. However, there are some limitations to this study that could be addressed in future work. Specifically, the dataset used in the study only included North American speakers, and the proposed approaches may not generalize well to people from different geographical regions.

In [4], the authors provide an overview of Deep Learning techniques and examine recent literature that employs these methods for speech-based emotion recognition. The review encompasses the databases used, emotions identified, and the contributions made towards speech emotion recognition, as well as the limitations associated with it. Deep Learning techniques employ key features in various applications, such as speech emotion recognition (SER) and natural language

processing (NLP). This allows for the learning of real-world data without the need for manual human labels. Traditional modeling techniques require a larger dataset to achieve accuracy in emotion recognition, which can be time-consuming. In contrast, Deep Learning methods are composed of various non-linear components that perform computations in parallel, but they require deeper layered architectures to be structured effectively.

In [5], the authors examine recent and relevant literature related to the various design components and methodologies of speech emotion recognition (SER) systems, thereby providing readers with a comprehensive understanding of the latest research on this topic. Moreover, while analyzing the current state of knowledge on SER systems, the authors highlight the need for further research to address the existing research gaps, which could be investigated by other researchers, institutions, and regulatory bodies. One of the primary challenges in SER systems is the difficulty in precisely defining the meaning of emotions, which can be complex and challenging to comprehend. The lack of agreement on the definition of emotions is reflected in the collection of databases. Another challenge is in reducing dimensionality and selecting appropriate features for analysis.

In[6],authors have focused on work dealing with the processing of acoustic clues from speech to recognize the speaker's emotions. The task of speech emotion recognition (SER) is traditionally divided into two main parts: feature extraction and classification. During the feature extraction stage, a speech signal is converted to numerical values using various front-end signal processing techniques. Extracted feature vectors have a compact form and ideally should capture essential information from the signal. In the back-end, an appropriate classifier is selected according to the task to be performed. the attention mechanism can improve the performance of the SER systems; however, its benefit is not always evident. Although AM modules have become a natural part of today's SER systems, AM is not an indispensable element for the achievement of high accuracies or even state-of-the-art results.

III. MODEL ARCHITECTURE

A) Layered Architecture

In a layered architecture for speech emotion recognition using a Multilayer Perceptron (MLP), the input layer of the MLP would take in acoustic features extracted from speech signals, such as MFCC coefficients or other features. The hidden layers of the MLP would learn representations of the input features through a set of weighted connections, and the

output layer would predict the emotional state of the speaker, such as happiness, sadness, or anger.

The number of neurons in the input layer would be determined by the number of acoustic features used as inputs. The number of hidden layers and the number of neurons in each hidden layer can be determined through experimentation and tuning. During training, the MLP would be presented with labeled speech data, where the emotion label is associated with each input speech sample. The network parameters can be learned through back propagation, where the difference between the predicted and actual emotions is used to update the weights and biases of the network.

In testing, the MLP would be presented with new speech data, and the output layer would predict the emotional state of the speaker. The performance of the MLP can be evaluated using metrics such as accuracy, precision, and recall.

B) Multilayer Neural Network

A multilayer neural network, such as a feed forward neural network or a recurrent neural network, can be used in speech emotion recognition. The input layer of the neural network would take in the speech signal, which could be represented as MFCC coefficients or other acoustic features. The hidden layers of the network would learn representations of the input features, and the output layer would predict the emotional state of the speaker, such as happiness, sadness, or anger. The neural network could be trained using labeled speech data, where the emotion label is associated with each input speech sample. The network parameters can be learned through back propagation, where the difference between the predicted and actual emotions is used to update the weights and biases of the network. The performance of the neural network can be evaluated using metrics such as accuracy, precision, and recall.

C) Multilayer Perceptron (MLP)

In the context of speech emotion recognition, MP could refer to several different things, so we will provide information on the two most common meanings:

Mel-frequency cepstral coefficients (MFCC) - MFCC is a feature extraction technique commonly used in speech emotion recognition. It involves computing the logarithm of the power spectrum of a speech signal, filtering the resulting spectrum with a bank of triangular filters spaced on the mel scale, and then computing the discrete cosine transform (DCT) of the logarithmic filter bank outputs. The resulting coefficients are called Mel-frequency cepstral coefficients (MFCC) and are used as input features for machine learning models such as neural networks. Max pooling - Max pooling

is a pooling operation commonly used in convolutional neural networks (CNNs) for feature extraction. It involves dividing the input feature map into non-overlapping rectangular regions and then taking the maximum value within each region. Max pooling can be applied to the output of a convolutional layer in a CNN, and the resulting pooled features can be fed into a multilayer neural network for emotion recognition. Overall, both MFCC and max pooling can be used in speech emotion recognition systems as part of the feature extraction process or as a part of the neural network architecture, respectively.

IV. PROPOSED SYSTEM

A) Dataset Description

The research work utilized the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) as its dataset. RAVDESS comprises 24 actors, each with 60 trials, and encompasses eight distinct emotions, including calm, happy, sad, angry, fearful, surprise, and disgust. The dataset contains a total of 1440 data samples.

B) Approach I

Two feature extraction approaches were proposed in the current study. The first method involved working directly on the audio dataset to derive numerical features, and this section provides additional information on the approach.

a) Feature Extraction

In order to predict the emotional content of a given speech, it is necessary to identify and extract meaningful features from the audio dataset. To achieve this, a combination of MFCC, Log Mel-Spectrogram, Chroma, Spectral centroid, and Spectral roll off features have been extracted using the librosa library. Following feature extraction, each file's features and labels have been transformed into a 2-D feature vector.

b) Dimensionality Reduction and Pre-processing

From the audio files, a total of 180 features have been extracted. To address the high dimensionality and sparsity of the dataset, additional pre-processing of the data was performed. The data was first normalized using the MinMaxScaler function from the sklearn library. Moreover, in order to reduce the dimensionality of the data and address issues of overfitting, PCA was employed. By eliminating highly correlated variables, PCA significantly reduced over fitting. A total of 80 important features were then selected using PCA, allowing for effective training and testing.

C) Approach II

In this section, the second approach for feature extraction is described. In this approach, we have utilized spectrograms as image features.

a) Feature Extraction

Another feature extraction method was utilized in this approach through transfer learning. Log Mel-Spectrogram images were generated from the input audio dataset, with the Mel-Spectrogram image for each file extracted and saved using the librosa library to its respective emotion class. Overall, a total of 1440 and 2800 images were extracted for the RAVDESS and TESS datasets, respectively. Proposed system is shown in figure 1.

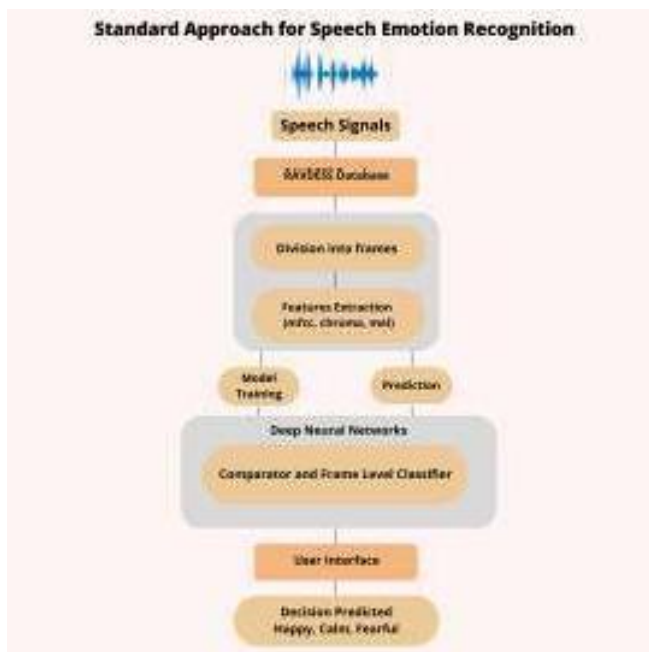


Figure 1: Proposed system for recognizing speech emotions

V. CONCLUSION AND FUTURE SCOPE

In conclusion, the proposed speech emotion recognition system based on MLP provides a promising solution to the challenges faced in customer support. The system has the potential to significantly improve the accuracy of emotion recognition and facilitate more effective customer support. The results of this research will contribute to the growing body of knowledge in the field of speech emotion recognition and provide a foundation for further advancements in this area. As a future system, we plan to add additional feature of Facial Recognition in real time to enhance the accuracy of the output.

REFERENCES

- [1] Bouzakraoui, Moulay & Sadiq, Abdelalim & Youssfi Alaoui, Abdessamad. (2020). Customer Satisfaction Recognition Based on Facial Expression and Machine Learning Techniques. *Advances in Science, Technology and Engineering Systems Journal*. 5. 594-594. 10.25046/aj050470.
- [2] X. Li and R. Lin, "Speech Emotion Recognition for Power Customer Service," 2021 7th International Conference on Computer and Communications (ICCC), Chengdu, China, 2021, pp. 514-518, doi: 10.1109/ICCC54389.2021.9674619.
- [3] Umer, M., Imtiaz, Z., Ullah, S., Mehmood, A., Choi, G. S., & On, B. W. (2020). Fake news stance detection using deep learning architecture (cnn-lstm). *IEEE Access*, 8, 156695-156706.
- [4] Aggarwal, A.; Srivastava, A.; Agarwal, A.; Chahal, N.; Singh, D.; Alnuaim, A.A.; Alhadlaq, A.; Lee, H.-N. Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning. *Sensors* 2022, 22, 2378.
- [5] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in *IEEE Access*, vol. 7, pp. 117327- 117345, 2019, doi: 10.1109/ACCESS.2019.2936124.
- [6] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," in *IEEE Access*, vol. 9, pp. 47795-47814, 2021, doi: 10.1109/ACCESS.2021.3068045.
- [7] Lieskovská, E.; Jakubec, M.; Jarina, R.; Chmulík, M. A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism. *Electronics* 2021, 10, 1163.
- [8] Mohanty, Subhadarshini; Mohapatra, Subasish; and Sahoo, Amlan (2022) "Speech Emotion Recognition System using Librosa for Better Customer Experience," *Graduate Research in Engineering and Technology (GRET): Vol. 1: Iss. 6, Article 7*.
- [9] Xu Dong An and Zhou Ruan 2021 *J. Phys.: Conf. Ser.* 1861 012064 "Speech Emotion Recognition algorithm based on deep learning algorithm fusion of temporal and spatial features".
- [10] Lausen, A., Hammerschmidt, K. Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. *Humanit Soc Sci Commun* 7, 2 (2020).
- [11] Poojary, Nagaraja & S, Dr & B.H, Akshath. (2021). Speech Emotion Recognition Using MLP Classifier. *International Journal of Scientific Research in Science and Technology*. 218-222. 10.32628/CSEIT217446.

Citation of this Article:

Dr. Rachna Somkunwar, Mr. Anil Kumar Gupta, Ishika Jain, Neha Shilvant, Rashi Pandey, Shreyas Deshpande, “Application Based System for Recognizing Speech Emotions” in proceeding of International Conference of Recent Trends in Engineering & Technology ICRTET - 2023, Organized by SCOE, Sudumbare, Pune, India, Published in IRJIET, Volume 7, Special issue of ICRTET-2023, pp 116-120, June 2023.
