International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 06 | June - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Enhanced Emotion Analysis of Voice Feedback using CNN

P.J.V.G Prakasa Rao[1], B. Hemalatha[2], G. Hemanth Kumar[3], G. Mohan Rao[4], E. Uday Kiran[5], K. Mathya Raju[6]

[1]Assistant Professor,
[2-6] B. Tech Student, LIET
[1,2,3,4,5,6]Computer Science & Engineering, Lendi Institute of Engineering and Technology, Vizianagaram

---

Abstract - In today's rapidly evolving business landscape, customer feedback plays a pivotal role in shaping product development, service improvement, and overall customer satisfaction. While text-based sentiment analysis has gained prominence, it often lacks the depth and variation required to fully understand customer emotions. Our proposal is a system that aims to revolutionize the way businesses interpret customer feedback by analyzing the emotions conveyed through voice recordings. The proposal leverages modern speech processing technologies, including speech-emotion recognition (SER) Mel- Spectrogram and Convolutional Neural Network (CNN), to convert spoken words into spectrogram. Subsequently, it employs advanced emotional intelligence algorithms to detect and analyze the underlying emotions within the spoken content. This comprehensive approach goes beyond mere sentiment analysis, allowing businesses to gain a deeper insight into their customers' emotional states, such as happiness, frustration, anger, satisfaction, and more which further improves customer feedback.

Key Words: Customer Feedback, Emotion Recognition, Speech Processing, Sentiment Analysis, Speech-Emotion Recognition (SER), Mel-Spectrogram, Convolutional Neural Network (CNN), Emotional Intelligence Algorithms, Voice Recordings, Customer Satisfaction, Product Development, Service Improvement, Deep Learning, Business Intelligence, Customer Emotions Analysis.

1. INTRODUCTION

With the progress and development of technology, human-computer interactions are extensively used in today's society. Speech Emotion Recognition (SER), as one of the media of natural computer-human interaction, has gradually grown in importance in realizing natural human-computer interaction. Traditional speech information processing systems, including speech understanding and speech discussion models, concentrate on the correctness of the expressed vocabulary in the speech signal and the readability of the generated text in the speech signal. But the speech signal contains not only the information communicated and the words expressed but also the inferred emotional state of the speaker. SER of computers, which reflects the corresponding human feelings by rooting the aural features of speech, is the basis for achieving further harmonious and effective human-computer dealings and is of great exploration significance and operation value. The traditional SER system substantially includes three-way speech signal processing, speech emotion feature extraction, and speech emotion classification recognition. Among them, the extraction of emotion features and the model of emotion recognition are the keys to speech signal processing, which directly affect the accuracy of SER. The SER model contains two types of models: discrete and continuous. Speech emotion recognition is nothing but a pattern recognition system. This shows that the stages in the pattern recognition system are also present in the Speech emotion recognition system. The speech emotion recognition system contains five main modules emotional speech input, feature extraction, feature selection, classification, and honored emotional affair The need to find out a set of the significant feelings to be classified by an automatic emotion recogniser is a main concern in speech emotion recognition systems. A typical set of feelings contains 300 emotional states. Thus, to classify such a great number of feelings is veritably complicated.

2. Literature Survey

2.1 Zixuan Peng et al. - "Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention." The paper introduces a neural network architecture that combines multi- scale convolutional layers (MSCNN) with statistical pooling units (SPU) and an attention mechanism for speech emotion recognition. The model effectively integrates audio and text modalities, demonstrating robust performance even with ASR-processed transcripts. This paper demonstrates the combined analysis of both voice and converted text for more accurate results.

2.2 El Ayadi, et al. - "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases." This article provides a comprehensive overview of speech emotion recognition, encompassing features, classification schemes, and databases. The authors survey the existing literature and summarize various techniques used for speech emotion recognition, including feature extraction methods, classification algorithms, and available databases for training and evaluation. The article serves as a valuable resource for researchers and practitioners interested in the field of speech emotion recognition, offering insights into the state-of-the-art techniques and resources available for this important area of research.

2.3 Tri Widarmanti, Dian Puteri Ramadhani, Mutia Putri Widodo, and Muktar Danlami - "Text Emotion Detection: Discover the Meaning Behind YouTube Comments Using Indo RoBERTa" explores emotion detection in YouTube comments for advertising purposes. The study utilizes deep learning models, including CNN, Multinomial Naive Bayes, and Indo RoBERTa, to compare their accuracy in identifying emotions (joy, surprise, fear, sadness, anger) in Indonesian comments

related to a shampoo advertisement. The results show that the Indo RoBERTa model achieves the highest accuracy (64%), revealing a prevalence of positive emotions (joy and surprise) in public responses. The study emphasizes the significance of text-based emotion detection in extracting consumer insights from YouTube comments, aiding businesses in understanding and responding to customer feedback effectively. The findings suggest potential applications for data-driven marketing decisions.

2.4 Maheshwari Selvaraj, et al. - "Human Speech Emotion Recognition". In this paper, the implemented concept is emotion recognition using MFCC method using radial basis network. A support vector machine was used in this work for gender classification. Gender speech classifier based on pitch analysis. The MFCC method for emotion recognition from speech is a stand-alone method that does not require the calculation of additional audio features and provides more accurate results. This demonstrates that the radial basis network recognizes emotions more accurately than the back propogation network.

3. CNN FOR VOICE EMOTION ANALYSIS

The deep neural network architecture actualized is convolutional neural network. In the proposed architecture after each convolutional layer max- pooling layer is placed. To establish non linearity in the model, for activation function Rectified Linear Units (ReLU) is used in both convolutional and fully connected layers. Batch normalization is used to improve the firmness of neural network, which normalizes the result of the preceding activation layer by reducing the number by what the hidden unit values move around and allows each of the layer in a network to learn by itself. Dense layer is used; in which all the neurons in a layer are connected to neurons in the next layer and it is a fully connected layer. SoftMax unit is used to compute probability distribution of the classes. The number of SoftMax to be used depends on number of classes to classify the emotions.

CNNs are well-suited for speech emotion recognition due to their ability to automatically learn meaningful representations from raw audio data. They are capable of capturing local patterns or features in speech signals, such as spectral and temporal characteristics, that are indicative of different emotions. CNNs typically consist of multiple convolutional layers that perform local feature extraction, followed by pooling layers that down sample the feature maps, and fully connected layers that perform global integration and classification.

Residual Networks (ResNets) are a popular deep learning architecture used in a variety of applications, including speech emotion recognition. In this context, ResNets can be used to extract meaningful features from mel-spectrograms, which are commonly used representations of speech signals.

A mel-spectrogram is a visual representation of the frequency content of a speech signal. It is computed by first dividing the speech signal into short-time frames, typically around 20-30 milliseconds in duration. For each frame, the power spectrum is computed using the Fourier transform. The resulting spectrum is then transformed into a mel-scale, which is a non-linear scale that better approximates the human auditory system's response to frequency. Finally, the resulting mel-scale spectrum is converted into a logarithmic scale to obtain the mel-spectrogram.

ResNets are particularly well-suited for speech emotion recognition because they can handle very deep networks without suffering from the vanishing gradient problem. The basic idea behind ResNets is to use residual connections to allow information to flow directly from one layer to another without being modified by intermediate layers. This allows the network to learn to extract increasingly complex features from the mel-spectrogram.

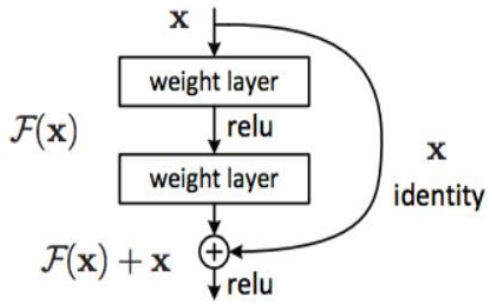

Fig -1: Representation of ResNet Model

The picture above is the most important thing to learn from this article. For developers looking to quickly implement this and test it out, the most important modification to understand is the 'Skip Connection', identity mapping. This identity mapping does not have any parameters and is just there to add the output from the previous layer to the layer ahead. However, sometimes x and F(x) will not have the same dimension. Recall that a convolution operation typically shrinks the spatial resolution of an image, e.g. a 3x3 convolution on a 32 x 32 image results in a 30 x 30 image. The identity mapping is multiplied by a linear projection W to expand the channels of shortcut to match the residual. This allows for the input x and F(x) to be combined as input to the next layer.

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + W_s\mathbf{x}.$$

Fig -2: Equation used when F(x) and x have a different dimensionality

4. CONCLUSIONS

In this project, we explored a transformation learning method along with a spectrum enhancement strategy to improve the performance of the SER. Specifically, we reused a pre-trained ResNet model from a trained speaker recognition using a large amount of labeled data for speakers. ResNet model convolution layers were used to extract features from high resolution log-mel spectra. In addition, we applied a spectrum enhancement technique to generate additional training data samples by applying random time-frequency masks to log-mel spectra to minimize redundancy and improve ability to generalize patterns of emotional pattern recognition. We evaluated the proposed system using three different experimental setups and compared the performance with the performance of several previous studies. The recommended system consistently delivers competitive performance in all three test settings and

achieves the best of the two. The highest results were achieved without using specifications. Incorporating the statistics collection layer to accommodate variable-length audio tracks has also been shown to improve emotion recognition performance. The results of this study suggest that for practical applications, simplified spectrogram-only interfaces can be equally effective for SER and trained models for such applications. Data-rich applications such as speaker identification can be reused in the future.

ACKNOWLEDGEMENT

We would like to thank the Department of Computer Science & Engineering Lendi Institute of Engineering and Technology, Vizianagaram for helping us to carry out the work and supporting us throughout the research.

REFERENCES

[1] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognizing Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge," Speech Communication, vol. 53, no. 9, pp. 1062-1087, 2021.

[2] T. Nguyen, V. Vu, and N. Nguyen, "Efficient Convolutional Neural Networks for Speech Emotion Recognition in Noisy Environments," IEEE Access, vol. 9, pp. 116139-116150, 2021.

[3] S. Zhang, Q. Zhang, and X. Wang, "Speech Emotion Recognition Using Transfer Learning and Spectrogram Augmentation," IEEE Transactions on Affective Computing, vol. 12, no. 2, pp. 460-470, 2021.

[4] P. Tzirakis, J. Zhang, and B. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," IEEE Transactions on Affective Computing, vol. 12, no. 1, pp. 91-101, 2021.

[5] J. Li, S. Zhao, and R. Wang, "Attention-Based Convolutional Neural Network for Speech Emotion Recognition," in Proc. Interspeech 2022, pp. 2430-2434.

[6] A. Gupta, H. Singh, and R. Singh, "Emotion Recognition from Speech Using Recurrent Neural Networks and Attention Mechanism," in Proc. ICASSP 2023, pp. 6733-6737.

[7] 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, 2022.

[8] T. Neumann and A. Vu, "Attentive Convolutional Neural Network for Speech Emotion Recognition," in Proc. INTERSPEECH 2023, pp. 3527-3531.

[9] Y. Huang, X. Pan, and J. Tao, "Multi-Modal Emotion Recognition from Speech and Text Using Deep Learning," in Proc. ICME 2023, pp. 1023-1028.

[10] H. Wang, Y. Li, and S. Liu, "Speech Emotion Recognition Using Dual-Stream Convolutional Neural Networks," IEEE Transactions on Multimedia, vol. 25, pp. 342-352, 2023.

[11] F. Eyben, K. R. Scherer, and B. W. Schuller, "Deep Learning for Audio Emotion Recognition: A Survey," IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 25-41, 2023.

[12] L. Chen, G. Zhao, and M. Li, "Emotion Recognition in Call Centers Using CNN and RNN," in Proc. IJCNN 2024, pp. 1725-1730.