Available *online* at : http://chipset.fti.unand.ac.id/

Journal on Computer Hardware, Signal Processing, Embedded System and Networking

| ISSN (Online) 2722-4422 |

Marathi Speech Emotion Recognition using Deep Learning Techniques

Vaibhav Narawade [1], Akhilesh Ketkar [2], Faizan Mulla [3], Madhur Nirmal [4], Divyansh Mishra [5]

[1] *Vaibhav Narawade, Dy Patil Deemed to be University, Ramrao Adik Institute of Technology, Navi Munbai – 400706, India*
[2] *Akhilesh Ketkar, Dy Patil Deemed to be University, Ramrao Adik Institute of Technology, Navi Munbai – 400706, India*

ARTICLE INFORMATION

Received: April 13th, 2023
Revised: February 21th, 2024
Available online: April 30th, 2024

KEYWORDS

Speech emotion recognition, Convolution neural network, Deep learning, Discrete-Time Fourier Transform

CORRESPONDENCE

E-mail: akh.ket.rt19@rait.ac.in

A B S T R A C T

In the project, an emotion recognition system from speech is proposed using deep learning. The goal of this project is to classify a speech signal into one of the five emotions listed below: anger, boredom, fear, happiness, and sadness. Snippets below from numerous Marathi movies and TV shows were used to construct the dataset for Marathi language samples which include 20 audio samples for anger, 19 for boredom, 5 for fear, and 11 for happiness. The proposed system first processes a speech signal from the time domain to the frequency domain using Discrete Time Fourier Transform (DTFT). Then, data augmentation is performed which includes noise injection, stretching, shifting, and pitch scaling of the speech signal. Next, feature extraction is performed in which 5 features were selected, which include Mel Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), Chroma STFT, Mel Spectrogram, and Root mean square value. These features were then fed to a Convolutional Neural Network (CNN). The efficiency of the suggested system employing the CNNs is supported by experimental findings. This model's accuracy on the test data is 80.33%, and its f1 values for anger, boredom, fear, happiness, and sadness are 0.85, 0.83, 0.50, 0.62, and 0.84, respectively.

INTRODUCTION

By identifying emotions from facial expressions, the essential framework of emotion identification research was established. Voice signals are recognized to carry significantly more information than text and are the fastest and most natural method of communication. Emotion recognition from speech has been studied for ages but finding the proper features that can give the highest accuracy is still a challenging task. Sentimental analysis is important to predict polarity orientation (positive/negative/ neutral conflict). Systems for tracking emotions can support subjective self-report measurement and help with emotion management [3]. The rapid increase in human-computer interaction and the rise in popularity of digital assistants like Alexa, Google Assistant, and Siri. Speech emotion recognition, which seeks to analyze the emotional state from speech signals, has gained more attention recently. Nonetheless, the issue of how to extract efficient emotional structure makes SER difficult work. The live emotion recognition module can present new deals and better opportunities to customers with immediate emotional feedback [2]. Our behavior and thoughts are highly affected by our emotions. The emotions you feel each day motivate you to take action and can cloud your judgments affecting large changes in your lifestyle. Emotion recognition provides an edge in many aspects of human life, also helping in the healthcare sector. Furthermore, it is essential to quickly and easily discern human emotions at a given time without directly addressing them. [7] The technology for recognizing emotions is far from perfect. Even though they can actually detect emotions, they nonetheless encounter and create problems. For 1 instance, a system may find subtle displays and emotions to be more frightening than overt ones [9]. Additionally, because it naturally associates specific facial expressions with particular emotions, it is unable to discriminate between genuine and false emotions and is susceptible to being tricked. While there has been significant progress in the field of emotion recognition, most of the existing research focuses on major languages like English, Mandarin, and Spanish. The recognition of emotions for regional languages, like Marathi, remains relatively unexplored. Marathi is spoken by over 83 million people worldwide, primarily in the Indian state of Maharashtra. Developing automated systems that can recognize and predict emotions accurately for Marathi could have significant implications for various fields, including education, healthcare, and entertainment. Predicting emotions using speech for Marathi is a challenging task due to the complexity of the language and the nuances in its pronunciation. Speech signals in

https://doi.org/10.25077/chipset.5.01.1-4.2024

Marathi have unique characteristics, and developing effective feature extraction techniques that can capture these characteristics is crucial for accurate emotion recognition. Moreover, speech recognition in Marathi is still in its infancy, with limited resources available for research and development. The goal of this study is to create a reliable Marathi-language system for predicting emotions using speech signals. This system could have numerous practical applications in various fields, including improving mental health diagnosis and treatment, enhancing language learning, and creating more engaging user experiences in entertainment. By addressing the challenges specific to Marathi speech signals, this research could contribute to the broader field of emotion recognition and provide valuable insights into speech processing for regional languages.

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x[n]\, e^{-jn\omega}$$

where j is the imaginary unit, ω is the frequency variable measured in radians/sample, and x[n] is the sequence to be transformed. The above equation expresses the DTFT as a continuous function of frequency ω, obtained by evaluating the sum of the sequence x[n] multiplied by complex exponential functions $e^{-j\omega n}$ for all values of n."

Database

The Marathi speech database used in this study is a newly created database specifically for this project. It consists of a corpus of 180 utterances of phrases. The speech data were collected from 6 speakers (3 males and 3 females). The speakers were native.

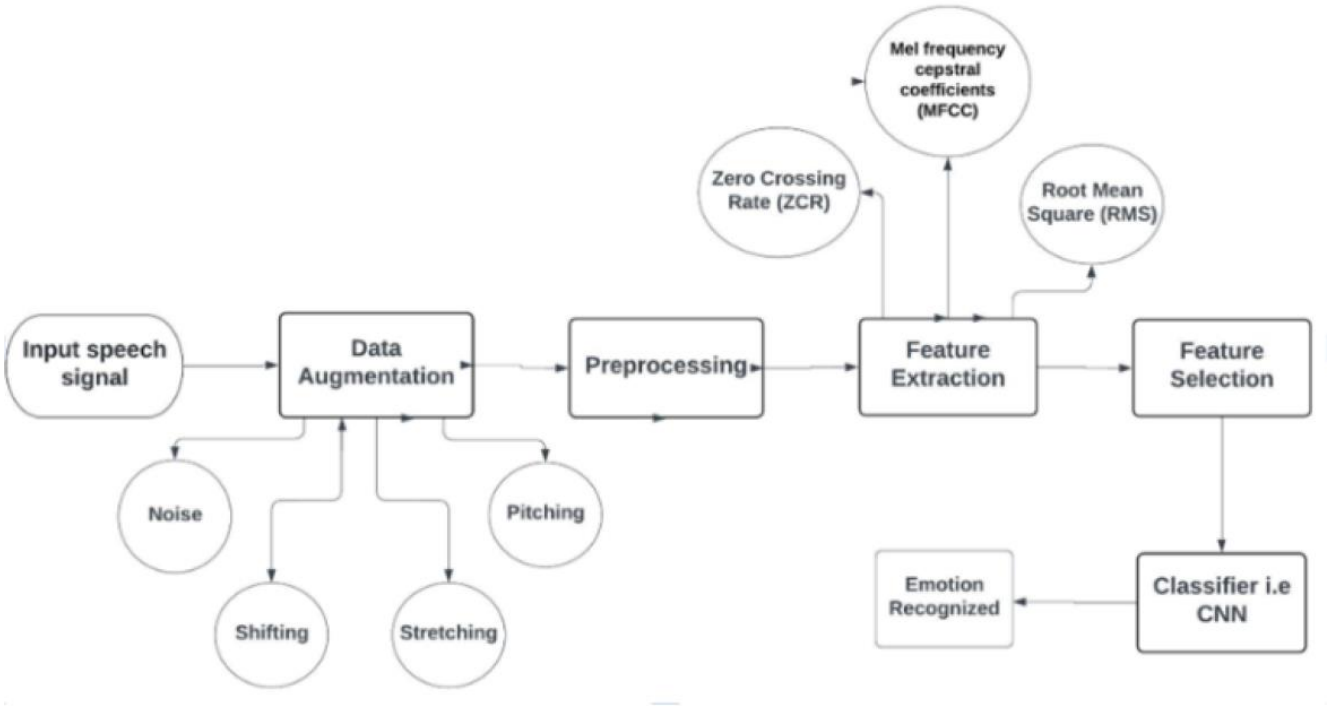

Figure 1. System Design

RESEARCH METHOD

Explaining research chronological, including research design, research procedure (in the form of algorithms, Pseudocode or other), how to test and data acquisition [1]-[3]. The description of the course of research should be supported references, so the explanation can be accepted scientifically [2], [4].

Objective Foundation

The proposed emotion recognition system for Marathi speech signals involves transforming speech signals from the time domain to the frequency domain using the Discrete Time Fourier Transform (DTFT). The DTFT is a mathematical tool used to analyze the frequency components of a discrete-time signal and is defined by the following equation:

Marathi speakers with a variety of ages, genders, and accents, to ensure a diverse and representative sample of the Marathi-speaking population. For each of the 5 emotions, each speaker was given a set of 5 sentences to utter.

The Marathi speech database was divided into training, validation, and test datasets. The training set consisted of [insert number of samples used for training] speech samples, the validation set consisted of [insert number of samples used for validation] speech samples, and the test set consisted of [insert number of samples used for testing] speech samples. The data division was done in a stratified manner to ensure that the distribution of different types of speech in the database was represented in each set.

The Marathi speech database created for this project is a valuable resource for speech recognition research in Marathi. It is publicly available for use by the research community and can be used as a benchmark for future speech recognition projects in Marathi.

https://doi.org/10.25077/chipset.5.01.1-4.2024

RESULTS AND ANALYSIS

Tables 1 and 2 exhibit the outcomes of the experiment on emotion recognition using the features.csv. Precision, recall, and the F1 score are included in the classification report for each emotion in Table 1. Recall measures the proportion of correctly recognized emotions among all actual feelings, while precision measures the proportion of correctly recognized emotions among all predicted emotions. The F1 score is calculated as the harmonic mean of precision and recall.

The results show that the system performed best in recognizing anger and sadness, with precision and recall scores of 0.79 and 0.92, and 0.79 and 0.88, respectively. The F1 score for anger and sadness was 0.85 and 0.84, respectively. The system performed less well in recognizing fear, with precision and recall scores of 1.00 and 0.33, and an F1 score of 0.50. The system performed even worse in recognizing boredom and happiness, with precision and recall scores of 0.83 and 0.83, and 0.80 and 0.50, respectively. The F1 scores for boredom and happiness were 0.83 and 0.62, respectively.

Table 2 presents the confusion matrix for the recognition experiment. The matrix shows that the system correctly recognized 11 instances of anger, 10 instances of boredom, 1 instance of fear, 4 instances of happiness, and 23 instances of sadness. However, the system made some incorrect predictions, with 2 instances of happiness being misclassified as anger, 1 instance of sadness being misclassified as fear, 2 instances of boredom being misclassified as sadness, and 1 instance of anger being misclassified as sadness.

Overall, the results show that the system performed well in recognizing anger and sadness but struggled with recognizing boredom and happiness. Further improvements in the feature extraction and classification methods are needed to improve the performance in recognizing these emotions.

Table 1. Classification Report for each of the emotions

Mood	Precision	Recall	F1 Score
Anger	0.79	0.92	0.85
Boredom	0.83	0.83	0.83
Fear	1.00	0.33	0.50
Happy	0.80	0.50	0.62
Sad	0.79	0.88	0.84

Table 2. Confusion matrix for recognition using the features.csv

	Anger	Boredom	Fear	Happy	Sad
Anger	11	0	0	0	1
Boredom	0	10	0	0	2
Fear	0	1	1	0	1
Happy	2	0	0	4	2
Sad	1	1	1	1	23

CONCLUSIONS

The experiment aimed at evaluating the performance of a speech recognition system for emotions using features extracted from Marathi speech signals. The results, shown in Tables 1 and 2, indicate that the system performed well in recognizing anger and sadness, but had some difficulty in recognizing boredom and happiness. The precision and recall scores for anger and sadness were high, while those for boredom and happiness were lower. To enhance the performance of the system, future work should focus on improving the feature extraction and classification methods. The use of deep learning techniques, such as CNNs and RNNs, may prove useful in this regard. Incorporating additional information, such as prosodic features, into the recognition process could also lead to improved performance. In conclusion, the results of this experiment provide a valuable foundation for future research into speech recognition systems for emotions in Marathi. The study highlights the need for further advancements in this field and the potential impact that these systems can have on various applications, including human-computer interaction and mental health assessment.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to everyone who contributed to the success of this research. Firstly, we would like to thank our guide Dr. Vaibhav Narawade for their guidance and support throughout this project. We are also grateful to Dr. Dhananjay Dhakane for their mentorship towards the project, invaluable insights and assistance with the data collection and analysis. Additionally, we would like to thank the participants who generously provided their speech samples for this study.

REFERENCES

[1] 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.

[2] W. Lim, D. Jang and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016, pp. 1-4, doi: 10.1109/APSIPA.2016.7820699.

[3] Yoon, WJ., Park, KS. (2007). A Study of Emotion Recognition and Its Applications. In: Torra, V., Narukawa, Y., Yoshida, Y. (eds) Modeling Decisions for Artificial Intelligence. MDAI 2007. Lecture Notes in Computer Science(), vol 4617. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-73729-2_43.

[4] M. S. Akhtar, A. Ekbal and E. Cambria, "How Intense Are You? Predicting Intensities of Emotions and Sentiments using Stacked Ensemble [Application Notes]," in IEEE Computational Intelligence Magazine, vol. 15, no. 1, pp. 64-75, Feb. 2020, doi: 10.1109/MCI.2019.2954667.

[5] M.Shamim Hossain , Ghulam Muhammad , Emotion Recognition Using Deep Learning Approach from Audio-Visual Emotional Big Data, Information Fusion (2018), doi:https://doi.org/10.1016/j.inffus.2018.09.008.

[6] K. -Y. Huang, C. -H. Wu, M. -H. Su and H. -C. Fu, "Mood detection from daily conversational speech using denoising autoencoder and LSTM," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 5125-5129, doi: 10.1109/ICASSP.2017.7953133.

[7] E. Lieskovska, M. Jakubec and R. Jarina, "Speech Emotion Recognition Overview and Experimental Results," 2020 18th International Conference on Emerging eLearning Technologies and Applications (ICETA), 2020, pp. 388-393, doi: 10.1109/ICETA51985.2020.9379218..

[8] Araño, K.A., Gloor, P., Orsenigo, C. et al. When Old Meets New: Emotion Recognition from Speech Signals. Cogn Comput 13, 771–783 (2021). https://doi.org/10.1007/s12559-021-09865-2.

[9] P. Tzirakis, J. Zhang and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5089-5093, doi: 10.1109/ICASSP.2018.8462677.

[10] 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.

[11] Liu, M. English speech emotion recognition method based on speech recognition. Int J Speech Technol 25, 391–398 (2022). https://doi.org/10.1007/s10772-021-09955-4.

NOMENCLATURE

ω	frequency variable

 https://doi.org/10.25077/chipset.5.01.1-4.2024