STM JOURNALS

Journal of
Polymer & Composites

ISSN: 2321-2810 (Online)
ISSN: 2321-8525 (Print)
Volume 12, Special Issue 5, 2024
DOI (Journal): 10.37591/JoPC

http://engineeringjournals.stmjournals.in/index.php/JoPC/index

Research JoPC

Development of Polymer-Based Sensors for Speech Emotion Recognition

Princy Tyagi*

Abstract

Traditional SER research often utilizes microphones with polymer components like Diaphragms and Membranes. Within some microphone designs, polymer membranes which plays a crucial role in converting sound pressure into electrical signals. The paper highlights the application (speech emotion recognition) and have tried to find polymer-based sensors. This work further delves deeper, investigating the performance of the CatBoost algorithm for emotion recognition in voice assistants designed for Indian languages. The research employs a dataset of labelled audio recordings encompassing various emotions in different Indian languages. Mel-Frequency Cepstral Coefficients (MFCC) and pitch features are extracted from the speech data. The CatBoost algorithm is then utilized for classification and compared to other commonly used algorithms like CNN, LSTM, XGBoost, and LightGBM. Our findings demonstrate that CatBoost achieves superior accuracy in emotion recognition compared to the other evaluated algorithms, particularly when using MFCC and pitch features. This highlights the potential of CatBoost for developing robust and efficient SER systems tailored for Indian language voice assistants. This research offers a novel perspective by bridging the gap between machine learning for SER and its potential application in voice assistants designed for specific languages. By leveraging CatBoost's capabilities, future voice assistants could better understand and respond to the emotional context of user interactions, potentially enhancing the user experience for Indian language speakers.

Keywords: Polymer components, Pattern recognition, Speech Emotion Recognition, Voice Analysis, Machine learning Algorithm.

INTRODUCTION

The vocalized form of human interaction is speech, which involves phonetically combining a specific number of vowel and consonant speech sound components to form each spoken word. Human languages come in tens of thousands of different variants, many of which are mutually unintelligible due to variations in vocabulary, grammatical organization, and speech sound unit collection. The majority of polyglots, or speakers of several human languages, are able to converse in two or more of them. Humans can sing because they possess the same vocal skills that allow them to speak.

***Author for Correspondence**
Princy Tyagi

Computer Science & Engineering, Swami Rama Himalyan University, Dehradun, Uttarakhand, India

Received Date: March 01, 2024
Accepted Date: July 16, 2024
Published Date: July 30, 2024

Citation: Princy Tyagi. Development of Polymer-Based Sensors for Speech Emotion Recognition. Journal of Polymer & Composites. 2024; 12(Special Issue 5): S268–S274.

In numerous civilizations, a written language has developed from speech, often exhibiting differences corresponding spoken language counterpart in terms of lexicon, syntax, and pronunciation. This phenomenon is known as diglossia. Speech serves as an internal soliloquy to reinforce and organize cognition, according to certain psychologists like Vygotsky, in addition to its role as a means of communication. In communication, which involves conveying

messages, speech is predominantly utilized. Information content in a message, represented by discrete symbols, and the rate of data transfer are measured in bits and bits per second (bps).

The process of identifying human emotions is known as emotion recognition. The accuracy with which individuals can gauge the emotions of others varies greatly. Technology's application to assist with emotional identification is a relatively recent field of research. The strategy often works best when it is used in a setting with a variety of modality options. The automation of physiological measurements. The retrieval of verbal cues from sound, sentiments from written content, and facial reactions from video using wearable technology has garnered significant interest.

In this work, we investigated the most effective method for detecting the emotional content of voice signals using a variety of AI architectures and 3D image models. After extensive testing, we advised using this method for Speech Emotion Recognition since it guarantees superior output with an increased accuracy.

LITERATURE REVIEW

Mutual expertise sharing and expertise exchange are intricately connected. An ordinary interaction between two individuals begins with shared identity and concludes with mutual trust. Before transferring human skills to robots, understanding people's emotion recognition is crucial. Over the years, considerable effort has been devoted to accurately isolating and identifying the relevant vocal attributes [1–4] Thanks to technology which can now recognize human speech and grasp commands. Although these programs are bilingual and excellent at obeying spoken instructions, they are unable to recognize emotions. When you type "I am in pain" into one of these programs, they will return music, movies, or articles. These uses would be significantly enhanced if they could perceive the speaker's emotion and perform searches accordingly [5].

Natural Language Processing (NLP) employs emotion detection to aid firms in formulating their marketing strategies. Rather than textual messages, individuals often share audio messages on social media platforms. However, as we must initially convert the voice to text before employing NLP, existing NLP techniques struggle when audio conversations are treated as text [6]. Voice emotion detection finds applications in medical technologies, Virtual reality is utilized in education, business development, and entertainment [7]. Preprocessing speech data to eliminate noise is essential before extracting the fundamental voice features [8]. After converting the speech waveform into a parametric representation, features can be extracted for further processing and analysis at a reduced data rate Effective feature extraction facilitates straightforward data categorization [9].

The algorithm must mimic human acoustics to extract speech features that resemble what people hear [10]. The primary properties of human hearing and speech are provided by a number of approaches, including "PLP, MFCC, LPCC, LPC, and LSF" [11]. To maintain the soundwise relevant aspects of the voice stream, MFCC linearly filters frequencies at low frequencies and logarithmically at high frequencies in order to retain key information. LSF evaluates the structure of the nasopharyngeal and oropharyngeal passages to linearly predict the physiological relevance of the image, while LPC assists in identifying formants and peaks in the spectrum [12]. The PLP method employs critical bands, intensity-to-loudness compression, and equal loudness pre-emphasis techniques to retrive information from speech, while LPCC leverages vocal tract characteristics to capture data related to various emotions [13]. Additional feature extraction techniques include discrete wavelets, integrated models, and deep learning. The WT extension of the DWT enables the simultaneous analysis of both the temporal and frequency domains of underlying signals for data extraction. Deep learning methods facilitate the utilization of I-vector and X-vector features [14, 15].

METHOD

Experimental Arrangement

The method we adopted for the present research adopted the experimental arrangement using the data set as shown in Table 1.

Table 1. Union of Tamil Nadu Universities-Southern National Education Statistics Consortium Data.

Emotion	Audio clips	Participation
Unbiased	140	15.35%
Terror	160	143.27%
Irritation	100	23.51%
Sorrow	180	10.18%
Joy	120	20.43%
Surprise	160	14.27%

Model

Further model has been proposed which is shown in Figure 1(a) and 1(b).

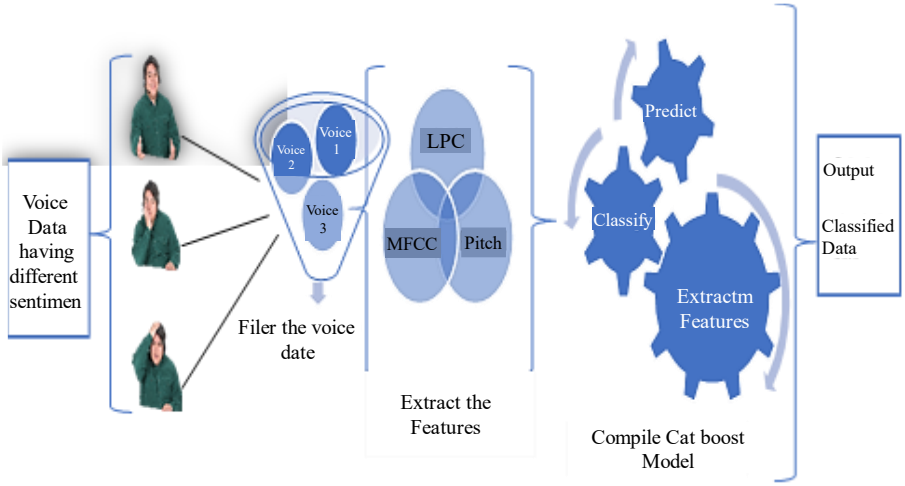

Figure 1. (a) Proposed Classification Model.

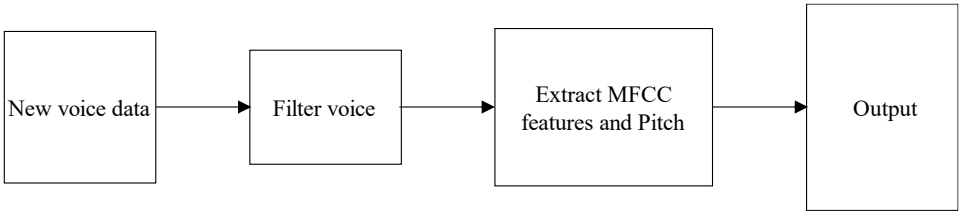

Figure 1. (b) Predictive Model.

Calculate Pitch

Then pitch calculator functions as per the Equation 1.

$$Pitch = \frac{Sampling\ Frequency}{max_l + index} \tag{1}$$

$$Here,$$
$$max_l : - maximum\ lag$$

RESULTS

The computational methods used to generate the conclusion are detailed alongside the outcomes.

A dialogue repetition graph from a Bollywood film is presented in Figure 2, created without the use of any refined. In Figure 3, the result of algorithm 2 is overlaid on the native speech. Figure 4, shows the pitch visualization plot.

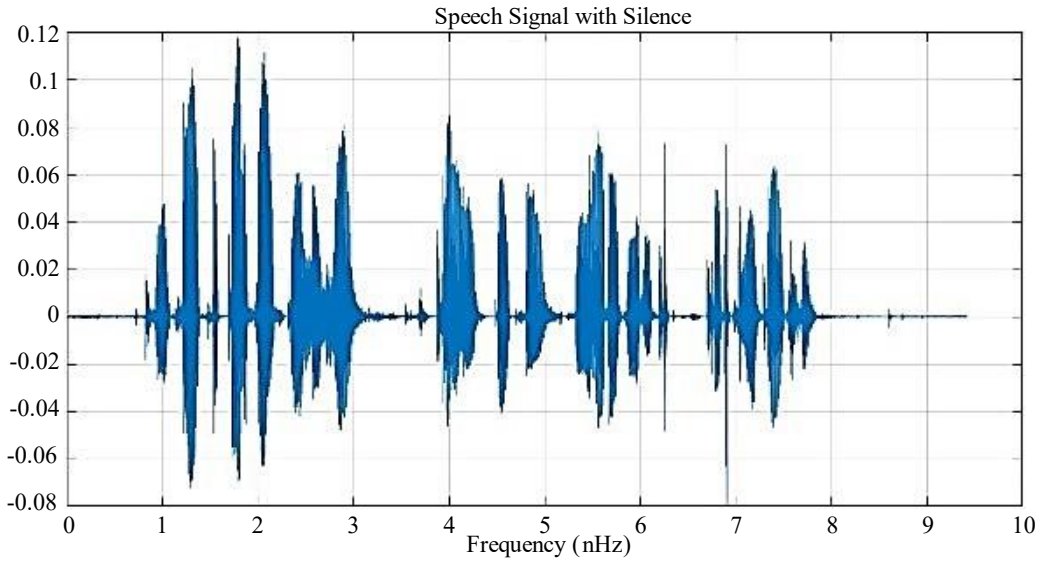

Figure 2. Native Speech (Created Without the Use of any Refined).

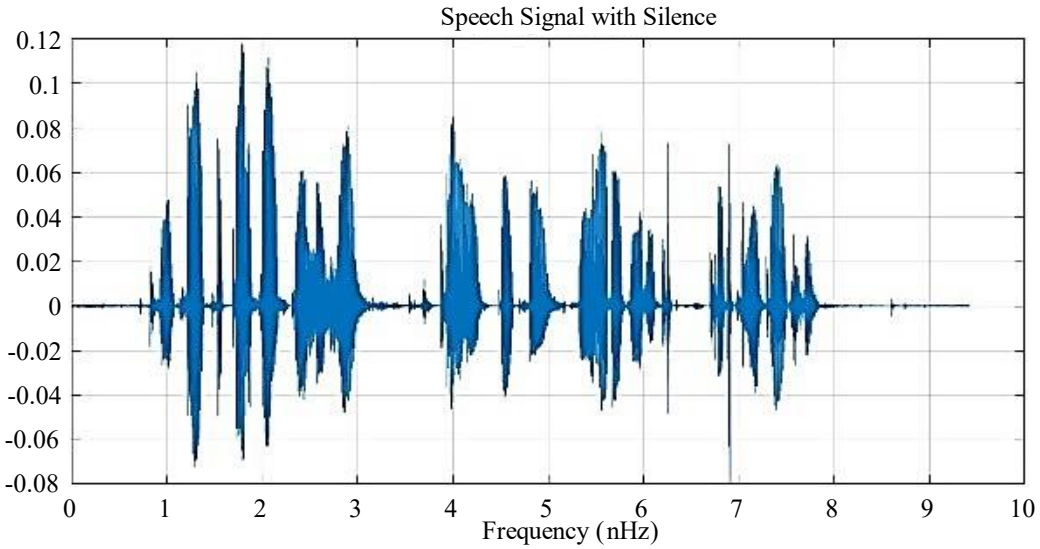

Figure 3. Refined Speech.

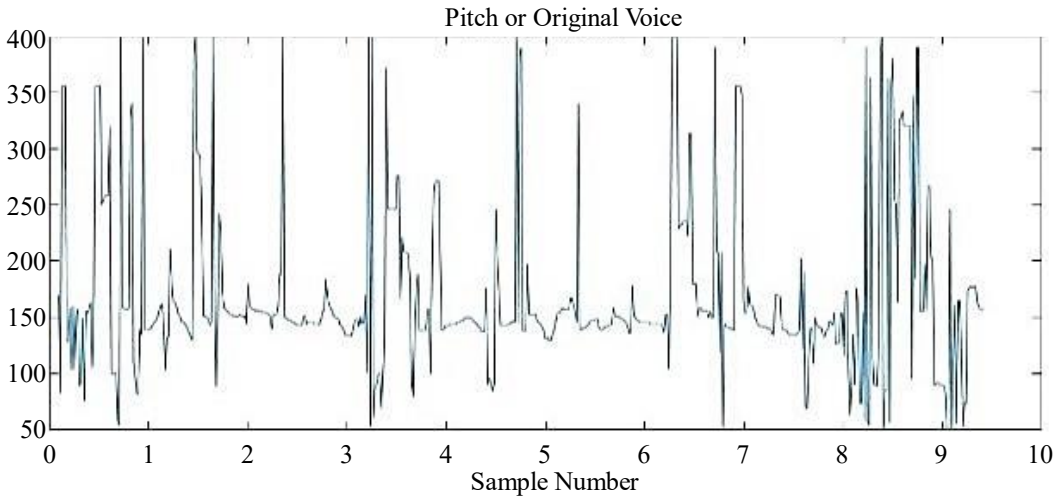

Figure 4. Pitch Visualization Plot.

If reproduced. Notably, there are minimal leading or trailing pauses observed in the recorded speech. The original signal is effectively suppressed during the intervals of silence in between. In Figure 5, the pitch variation of the filtered voice is depicted. It is apparent that the pitch of the filtered voice is affected by the frequency shift. Nonetheless, the voice remains unaltered post-filtering, with only the fast-forward segment of the video being played.

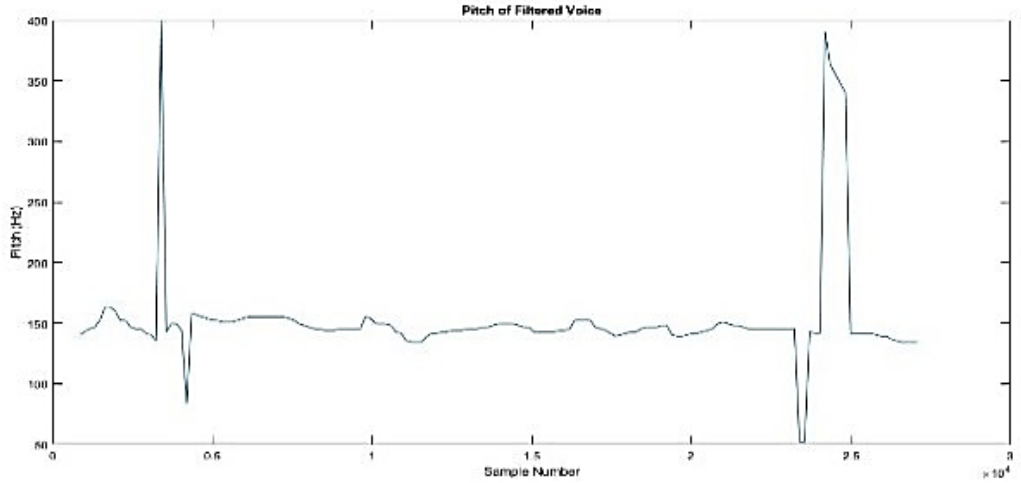

Figure 5. Impacts of Window Function Application(Pitch Variation of the Filtered Voice is Depicted)

CatBoost is employed in Table 2 for training and prediction purposes. In this methodology, 20% of the dataset is set aside for evaluation and confirmation purposes, with the left 80% designated for educating. The rationale for selecting CatBoost is elaborated upon in Tables 2 and 3. The results derived from these tables reinforced our choice of CatBoost for prediction and categorization assignments. Furthermore, we conducted additional experiments using other techniques such as CNN, LSTM, XGBoost, and LightGBM. The findings presented in Tables 2 and 3 indicate that CatBoost outperforms these alternatives, particularly when utilizing features such as MFCC, LPC, and Pitch. One key advantage of CatBoost is its ability to extract its own features from the data.

Table 2. Analysis of Sadness Utilizing Different Computational Methods

Algorithm	Precision	Recall	Accuracy
CNN	0.823	1.0	0.893
LSTM	0.842	1.0	0.902
XBoost	0.861	1.0	0.924
LightGBM	0.844	1.0	0.901
CatBoost	0.922	1.0	0.954

Table 3. Analysis of Happiness Utilizing Different Computational Methods.

Algorithm	Precision	Recall	Accuracy
CNN	0.81	1	0.88
LSTM	0.83	1	0.90
XBoost	0.85	1	0.91
LightGBM	0.83	1	0.90
CatBoost	0.90	1	0.95

CONCLUSION

After comparing every classification method, it was found that CatBoost provided the greatest accuracy. Indian languages are being addressed through CatBoost. The dataset offered has limitations

since there are so few speakers. Training and categorization on a little dataset are difficult. CatBoost demonstrates remarkable capability in extracting and identifying voices, even in scenarios akin to a cocktail party. Its accuracy in Indian languages, especially in cases with limited datasets, is notable, reaching up to 95.05 percent. Leveraging CatBoost, one could develop Indian-speaking voice bots that remain functional even amidst considerable noise. This technology could significantly simplify communication between native language speakers and computers. Businesses stand to benefit from CatBoost's ability to identify speakers' emotions and formulate appropriate responses. Additionally, online education providers could leverage CatBoost's insights to gauge students' emotional engagement with their programs via thermal imaging. This expertise could be instrumental in developing more effective lesson plans and teaching methodologies.

Future Scope

In the future, our goal is to incorporate more languages and sounds into our system. In scenarios where there's a cacophony of voices, we aim to explore the efficacy of having multiple natural speakers speaking simultaneously. Despite potential overlaps, our model should be flexible to incorporate additional Indian languages. This endeavor could prove beneficial for companies striving to develop technology capable of recognizing patients' emotions. Emotion recognition holds promise in aiding medical science to better assist patients

REFERENCES

1. Sharma U, Maheshkar S, Mishra AN. Study of robust feature extraction techniques for speech recognition system. In2015 International conference on futuristic trends on computational analysis and knowledge management (ABLAZE) 2015 Feb 25 (pp. 654-658). IEEE.
2. Gupta H, Gupta D. LPC and LPCC method of feature extraction in Speech Recognition System. In 2016 6th international conference-cloud system and big data engineering (confluence) 2016 Jan 14 (pp. 498-502). IEEE.
3. Chadha AN, Zaveri MA, Sarvaiya JN. Optimal feature extraction and selection techniques for speech processing: A review. In 2016 International Conference on Communication and Signal Processing (ICCSP) 2016 Apr 6 (pp. 1669-1673). IEEE.
4. Letaifa LB, Torres MI, Justo R. Adding dimensional features for emotion recognition on speech. In 2020 5th international conference on advanced technologies for signal and image processing (ATSIP) 2020 Sep 2 (pp. 1-6). IEEE.
5. Mehra P, Verma SK. Comparing Classifiers for Recognizing the Emotions by extracting the Spectral Features of Speech Using Machine Learning. In 2023 International Conference on Device Intelligence, Computing and Communication Technologies,(DICCT) 2023 Mar 17 (pp. 387-391). IEEE.
6. Lakomkin E, Zamani MA, Weber C, Magg S, Wermter S. Incorporating end-to-end speech recognition models for sentiment analysis. In 2019 International Conference on Robotics and Automation (ICRA) 2019 May 20 (pp. 7976-7982). IEEE.
7. Sajjad M, Kwon S. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE access. 2020 Apr 27;8:79861-75.
8. Garg K, Jain G. A comparative study of noise reduction techniques for automatic speech recognition systems. In 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI) 2016 Sep 21 (pp. 2098-2103). IEEE.
9. Pardede HF. On noise robust feature for speech recognition based on power function family. In 2015 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) 2015 Nov 9 (pp. 386-390). IEEE.
10. Alim SA, Rashid NK. Some commonly used speech feature extraction algorithms. London, UK:: IntechOpen; 2018 Dec 12.
11. Gill AS. A review on feature extraction techniques for speech processing. International Journal Of Engineering and Computer Science. 2016 Oct;5(10):18551-6.
12. Hermansky H. Perceptual linear predictive (PLP) analysis of speech. the Journal of the Acoustical Society of America. 1990 Apr 1;87(4):1738-52.

13. Luo Y, Fu Q, Xie J, Qin Y, Wu G, Liu J, Jiang F, Cao Y, Ding X. EEG-based emotion classification using spiking neural networks. IEEE Access. 2020 Mar 4;8:46007-16.
14. Amini MM, Matrouf D. Data augmentation versus noise compensation for x-vector speaker recognition systems in noisy environments. In 2020 28th European Signal Processing Conference (EUSIPCO) 2021 Jan 18 (pp. 1-5). IEEE.
15. Wu J, Hua Y, Yang S, Qin H, Qin H. Speech enhancement using generative adversarial network by distilling knowledge from statistical method. Applied Sciences. 2019 Aug 18;9(16):3396.

