Journal of Advanced Zoology
ISSN: 0253-7214
Volume 44 Issue 05 Year 2023 Page 364:375

# Enhancing Emotion Classification in Malayalam Accented Speech: An In-Depth Clustering Approach

**Rizwana Kallooravi Thandil[1][0000-0001-9305-681X], Mohamed Basheer K.P[2] [0000-0001-9305-681X]**

*[1]*,[2] Sullamussalam Science College, Areekode1ktrizwana@gmail.com Kerala India*

***Corresponding Author:** Rizwana Kallooravi Thandil*
**Sullamussalam Science College, Areekode1ktrizwana@gmail.com Kerala India*



***Abstract***

Accurate emotion classification in accented speech for the Malayalam language poses a unique challenge in the realm of speech recognition. In this study, we explore the application of various clustering algorithms to this specific dataset, evaluating their effectiveness using the Silhouette Score as a measure of cluster quality. Our findings reveal significant insights into the performance of these algorithms. Among the clustering methods, Affinity Propagation emerged as the frontrunner, achieving the highest Silhouette Score of 0.5255. This result indicates a superior cluster quality characterized by well-defined and distinct groups. OPTICS and Mean Shift Clustering also demonstrated strong performance with scores of 0.4029 and 0.2511, respectively, indicating the presence of relatively distinct and well-formed clusters. In addition, we introduced Ensemble Clustering (Majority Voting), which achieved a score of 0.2399, indicating moderate cluster distinction. These findings provide a valuable perspective on the potential advantages of ensemble methods in this context. Our experiment results shed light on the effectiveness of various clustering methods in the context of emotion classification in accented Malayalam speech. This study contributes to the advancement of speech recognition technology and lays the groundwork for further research in this area..

***Keywords: Speech Emotion Recognition (SER), Accented Speech Recognition, Clustering Techniques, Speech Feature Engineering.***

## 1 Introduction

Speech Emotion Recognition (SER) has garnered significant attention both in academia and industry due to its broad applicability in fields such as human-computer interaction and affective computing. A key facet of efficient SER lies in feature engineering, which plays a pivotal role in enhancing classification accuracy. While extensive research efforts have been dedicated to this domain, there remain challenges pertaining to optimal speech feature selection and the correct application of feature engineering techniques. The ability to detect and interpret emotions from speech signals has evolved into a fundamental aspect of modern speech processing technology. Beyond the mere conveyance of spoken words, the underlying emotional cues hidden within the

vocal tones and patterns offer profound insights into a speaker's state of mind, enriching applications across diverse domains such as human-computer interaction, sentiment analysis, and customer service. However, the task of emotion detection in speech becomes notably intricate when considering the prevalence of accents. Accents, shaped by cultural and geographical influences, introduce a captivating layer of complexity, imbuing speech with unique tonal, rhythmic, and phonetic variations.

This study embarks on a journey at the intriguing intersection of emotion detection and accented speech, with a specific emphasis on the Malayalam language. The insights gleaned from our study hold promise for applications in voice assistants, emotion-aware computing, and cross-cultural communication, as they provide a deeper understanding of the nuances of emotional expression in accented speech. As we embark on this academic voyage, we recognize the importance of addressing the underexplored territory of accented speech emotion detection and its implications. By delving into this uncharted domain, we aspire to contribute not only to the realm of speech processing but also to the broader arena of cross-cultural communication and understanding. The objectives of this study can be listed as:

1) To Understand the Complexities of Accented Speech: Delve deep into the Malayalam language's accented speech signals to comprehend the inherent challenges and intricacies they present for emotion detection.
2) Development of a Robust Dataset: Compile and curate a comprehensive dataset that encompasses diverse emotional states, ensuring a wide representation of the Malayalam language's accents.
3) Innovative Feature Engineering: Design and implement novel feature engineering techniques that can effectively capture the unique emotional nuances present in accented speech signals.
4) Application of Clustering Techniques: Utilize advanced clustering methodologies to group similar emotional states from the dataset, facilitating more accurate and efficient emotion detection.
5) Evaluate the Efficacy of the Proposed Method: Conduct a series of experiments and analyses to gauge the performance of the developed feature extraction and clustering techniques, comparing them with existing methodologies, if any.
6) Contribute to the Field of Emotion Detection: By integrating unique feature extraction with clustering techniques, aim to set a new standard in the domain of emotion detection from accented speech, enriching the existing body of knowledge.

## 2 Literature Review

In the pursuit of advancing the field of emotion classification in accented speech for the Malayalam language, a thorough exploration of the existing body of knowledge becomes imperative. This literature review section is dedicated to elucidating the foundational concepts, methodologies, and significant research contributions that have paved the way for our investigation. The review is organized to guide readers through the intricate landscape of emotion classification in accented speech, establishing the context within which our study is situated. The recognition of emotions in spoken language is a multifaceted challenge with broad-reaching implications for various domains, including human-computer interaction, affective computing, and customer service. It transcends the realm of mere linguistic understanding, delving into the intricate world of paralinguistic cues. Emotions, conveyed through variations in pitch, tone, speech rate, and other acoustic features, add a nuanced layer of communication to speech. This enables machines to not only understand the content of spoken words but also interpret the speaker's emotional state. Such capabilities have the potential to revolutionize applications like chatbots that respond empathetically, voice assistants that adapt to users' emotional needs, and automated customer service systems that gauge customer satisfaction in real-time.

While the recognition of emotions in standard, unaccented speech has seen considerable progress, the introduction of accents adds an additional layer of complexity. Accents, reflective of the speaker's cultural and geographical background, introduce distinct phonetic variations that can significantly influence the acoustic features associated with emotions. These variations necessitate specialized approaches for accurate emotion classification in accented speech. This literature review aims to survey the landscape of research that has grappled with this very challenge, and to identify the gaps and opportunities that guide our current study. As we embark on this journey through the relevant literature, we will traverse the rich tapestry of emotion classification methodologies, the influence of accents on speech, and the specific nuances of emotion classification within the Malayalam language. Through this exploration, we aim to build a comprehensive foundation that underpins our research, providing a springboard for the innovative clustering techniques employed in our study.

The study presented by [1] introduces a novel feature optimization approach, employing a clustering-based genetic algorithm. The proposed methodology departs from the conventional random selection of new feature sets in genetic algorithms. Instead, it integrates clustering at the fitness evaluation level to identify outliers that should be excluded from the next generation of features. This innovative approach was rigorously tested using well-established speech emotion databases and the results obtained from the experiments conducted by [1] demonstrated the efficacy of the proposed technique in enhancing emotion classification in speech. For general speakers (comprising both male and female), the recognition rates were notably high, with 89.6% on EMO-DB, 86.2% on RAVDESS, and 77.7% on SAVEE in speaker-dependent experiments. In speaker-independent experiments, recognition rates of 77.5% on EMO-DB, 76.2% on RAVDESS, and 69.8% on SAVEE were achieved.

The paper [2] presents an emotion recognition system for improving human-robot interactions by analyzing speech signals. It employs a multi-step approach, involving the extraction of 88-dimensional audio features, generation of spectrograms, and the selection of keyframes through k-means clustering. These keyframes' corresponding spectrograms are organized into a 3D tensor, which is used to train a 3D Convolutional Neural Network (CNN). The results from experiments conducted on three databases demonstrate the system's superiority over existing methods, p Dutt and Gader (2023) introduced the WaDER method, which incorporates an autoencoder, 1D CNN, and LSTM networks for Speech Emotion Recognition (SER). Their approach yielded an Unweighted Accuracy (UA) of 81.45% and a Weighted Accuracy (WA) of 81.22% in speaker-dependent experiments [3]. Zisad and their team (2020) devised a system for recognizing emotions in individuals with neurological disorders. Their model utilizes tonal properties such as Mel-Frequency Cepstral Coefficients (MFCCs) and the RAVDESS audio speech and song database [4]. Abdelhamid et al. (2022) employed cascaded layers of feature learning blocks. Their method involves extracting high-level features from the log Mel-spectrum of speech samples, effectively capturing local correlations and contextual information, resulting in enhanced emotion recognition performance [5].

Lee and Kim (2020) conducted a study on speech emotion recognition, with a focus on deep learning techniques, including multi-layer perceptron (MLP) and convolutional neural networks (CNN). They used 1D data extracted from speech files and two-dimensional Mel-spectrogram images to train the models. The test accuracy reached approximately 60%, with the CNN model achieving the highest accuracy among the approaches examined [6]. Abdulmohsin et al. (2021) introduced a novel statistical feature extraction method for SER. This approach utilizes the variance of feature distribution around the mean, employing multiple degrees of standard deviation on both sides of the mean [7]. In a separate context, Sunny et al. (2012) developed an Automatic Speech Recognition (ASR) system for isolated spoken words in Malayalam. They utilized Discrete Wavelet Transforms (DWT) and Artificial Neural Networks (ANN). The method was evaluated with 50 speakers uttering 20 isolated words each, demonstrating high recognition accuracy. The combination of DWT and ANN proved effective in feature extraction from speech signals for automatic speech recognition, significantly reducing computational complexity and feature vector size. The result was an impressive recognition accuracy of 90% [8].

## 3 Data Collection

The primary data for this research was sourced from the publicly available YouTube platform. Given YouTube's vast collection of videos, it offers a diverse repository of speech samples, catering to numerous accents and emotional expressions. To maintain the research's focus on the Malayalam language and its varied accents, specific criteria were set for video selection: Videos must predominantly be in the Malayalam language. Preference was given to videos that showcased explicit emotional expressions, aligning with the seven emotions under consideration: anger, disgust, fear, happiness, neutral, sadness, and surprise. The dataset comprises a rich mix of speech samples representing the seven emotions from across different accented data. Table 1 shows the statistics of the accented emotional data collected for the study.

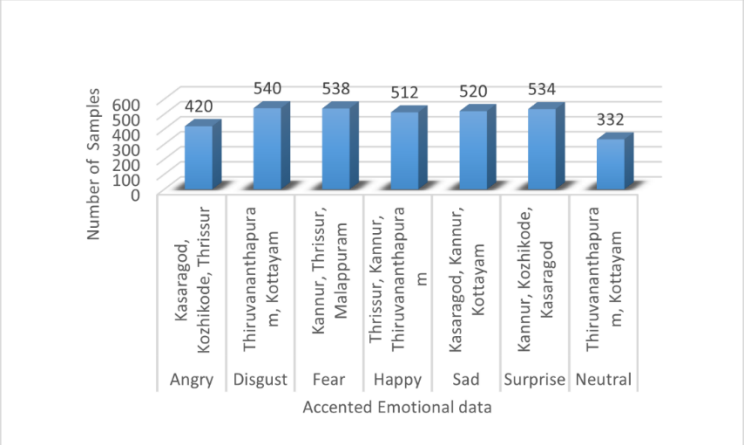


**Fig. 1.** Statistics of the Dataset

This distribution, as visualized in the accompanying chart, offers a balanced representation of emotions, ensuring that no emotion is underrepresented. The data collection and preparation steps are shown in Fig. 2.

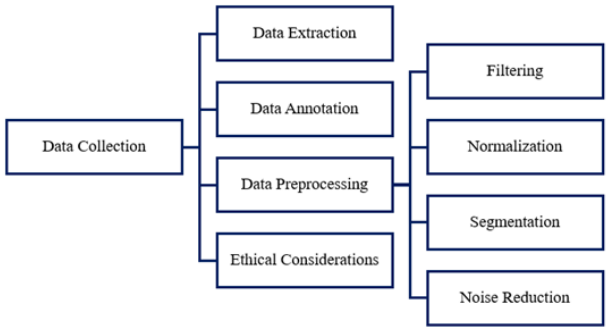


**Fig. 2.** Data Collection and Preparation Process

In the initial phase of our research, we identified suitable video content that would serve as the source for our dataset. To focus on the specific emotions of interest, we carefully extracted audio segments from these videos using advanced audio extraction tools like audacity. The data annotation part was one of the crucial steps in the dataset curation.

### 3.1 Data Preprocessing

The data preprocessing section involves four different approaches in the study- filtering, normalization, segmentation, and noise reduction.

**1. Filtering**

By applying this band-pass filter to each audio sample, the processing pipeline laid the foundation for cleaner and more intelligible speech signals. This enhanced audio quality is instrumental in subsequent stages of speech analysis, ensuring that the emotional nuances and features within the speech are captured with greater precision. Overall, the band-pass filtering step plays a pivotal role in preparing the audio data for accurate and insightful emotion recognition and analysis.

The Implementation Details

i. The Butterworth filter design was utilized, with user-specified order (default set to 5).
ii. A specific frequency range most relevant to the Malayalam language speech was defined: a low cutoff frequency of 75.0 Hz and a high cutoff frequency of 3400.0 Hz.

**2. Audio Normalization**

In the context of audio data processing, normalization stands as a critical step in the pursuit of equitable and accurate analysis. Its primary objective is to harmonize the amplitude levels across all audio samples, thereby avoiding any undue emphasis or suppression of certain samples based on their inherent amplitude characteristics. Initially the dataset is comprised of audio samples with varying volume levels. Without normalization,

these differences in amplitude can introduce biases during subsequent analysis, where louder samples may dominate, and quieter ones might be overshadowed. To counteract this, normalization is applied uniformly to all samples, ensuring that they are brought to a consistent amplitude level. This process can be likened to standardizing the playing field, where each audio sample is granted an equal opportunity to convey its inherent emotional nuances. By doing so, we mitigate the influence of volume disparities on the results, allowing the true emotional content of the speech to shine through.

Normalization, in essence, serves as an essential preparatory step to eliminate unintended bias and to promote fairness in emotion recognition and analysis. It ensures that the true emotional expressions within each speech sample are faithfully captured, free from the confounding impact of varying amplitudes. This step, although seemingly technical, is instrumental in preserving the integrity of the emotion recognition process and maintaining the equitable treatment of all audio samples.

The Implementation Details:

i. Audio normalization was achieved by dividing each audio sample by its maximum absolute value, ensur ing that the waveform's peak amplitude is consistent across all audio files.
ii. The normalized audio data was then multiplied by 32767 to bring it within the range of 16-bit PCM WAV format and saved as a new WAV file. The saved files were prefixed with 'normalized_' to indicate the applied transformation.
iii. In essence, through band-pass filtering, unwanted frequencies that might not contribute significantly to emotion detection were removed. Subsequent normalization ensured that every audio sample was on a level playing field, paving the way for robust and consistent analysis in the later stages of the research.

**3. Segmentation**

To facilitate consistent and manageable analysis, we segmented longer audio clips into uniform, shorter segments. This approach ensured that each segment received equal attention during the analysis, promoting a fair assessment of emotional expressions.
The Implementation Details:

i. Decide on the desired duration for each shorter segment, considering the nature of emotional data. Given that the long sentences are 10 seconds or longer, a shorter segment length that aligns with the typical duration of emotional expressions within those sentences shall be chosen.
ii. Create a loop or iteration process to go through each of the 10-second audio segments. This step prepares the segments for further division.
iii. Within the loop, segment each 10-second audio clip into shorter, uniform segments of the predetermined duration. These shorter segments could be 5 seconds each. Ensure that the segmentation process maintains consistency and that each shorter segment encapsulates distinct emotional expressions.

**4. Noise Reduction**

The clarity of speech samples is paramount in emotional analysis. Thus, we employed noise reduction techniques to eliminate any residual background noise, enhancing the overall quality of the audio data. This step aimed to enhance the precision of emotion recognition.

**5. Ethical Considerations:**

While our dataset was sourced from publicly available content, we were acutely aware of the need to uphold ethical standards and respect privacy concerns. To address this, we took measures to exclude any personal information or identifiers from the dataset. The use of the data was strictly confined to academic and research purposes, adhering to the terms of service provided by the platform from which the content was sourced (YouTube). These ethical considerations underscore our commitment to responsible data usage and research practices. The data preprocessing stage is pivotal in ensuring the quality and consistency of the audio data before subsequent analysis. The primary goal is to optimize the data to highlight the most relevant features while minimizing any unwanted noise or discrepancies. Two primary preprocessing steps were performed: band-pass filtering and audio normalization.

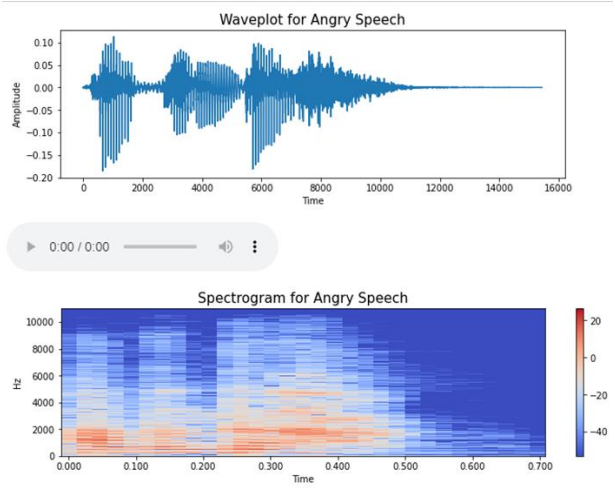


**Fig.2**. The Waveplot and Spectrogram of the Sample Emotion Used in the Study

## 4 Feature Engineering

Feature extraction is a crucial phase in speech processing, especially in emotion detection, where subtle audio characteristics can carry significant emotional cues. In this study, an extensive set of features was derived using a combination of techniques, ensuring a comprehensive representation of the audio data's emotional content.584 features were extracted in this study using different feature extraction techniques.

Before extraction, audio signals were padded to a consistent length, corresponding to a given FFT size, using the padding function. This step ensures that all audio samples have uniform dimensions, facilitating consistent feature extraction. The Spectral Contrast measures the difference in amplitude between peaks and valleys in the sound spectrum. The polyfeatures are used to calculate the polynomial approximations to the spectrogram data. The tempogram represents the rhythm patterns in the audio. Tonnetz calculates the tonal centroid features, providing insights into the harmonic relations in the audio data. And for Harmonic-to-Noise Ratio (HNR), a measure of the sound's periodicity was derived, indicating the ratio of the harmonics to the noise in the signal. Formant frequencies are used to extract specifically the first two formants (F1 and F2), were extracted. These frequencies are essential in characterizing speech and can carry significant emotional cues. Pitch Variability is the standard deviation of the fundamental frequency (F0) was computed, providing insights into the pitch variations in the audio. Mel-frequency cepstral coefficients (MFCCs), along with their first and second derivatives (delta and delta2), were computed. These coefficients capture the short-term power spectrum of sound and are widely used in speech and audio processing. Zero Crossing Rate (ZCR) measure indicates the rate at which the signal changes its sign, reflecting the noisiness or percussiveness of the audio. Chroma STFT relates to the twelve different pitch classes and is used to describe harmony. Root Mean Square Value (RMS) indicates the audio's energy, which can be a good proxy for loudness. Mel Spectrogram represents the short-term power spectrum of sound in the Mel scale.

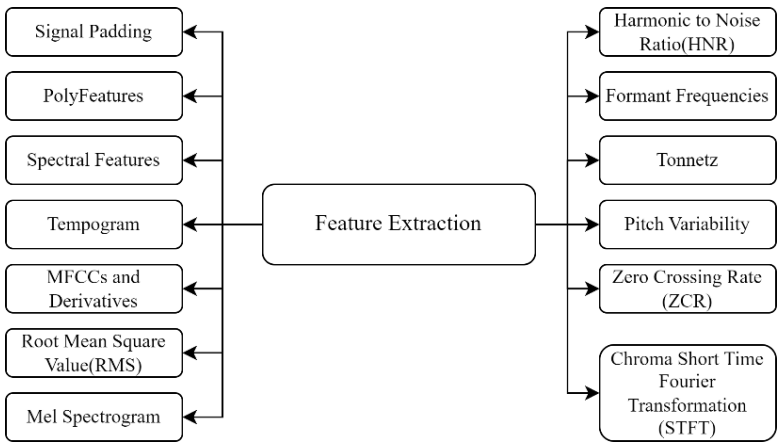


**Fig.3**. The Feature Extraction Techniques Used in the Study

**4.1Data Augmentation:**
To enhance the robustness of the features and improve the model's generalization capabilities, the following augmentations were applied to the audio data:
i. Noise Addition: White noise was introduced to the audio.
ii. Stretching and Pitching: The audio was stretched, and a subsequent pitch modification was applied.

**4.2Feature Reduction**
Feature reduction is an essential step in machine learning and deep learning applications, especially for high-dimensional data. By reducing the number of features, we can alleviate the curse of dimensionality, speed up model training, and potentially enhance model generalization by reducing the risk of overfitting. In our study on accented speech recognition for the Malayalam language, this reduction process was paramount given the complexity of audio features.

i. Feature Correlation Analysis:
Our initial step was to determine inter-feature correlations within our dataset. A heatmap was generated to visually assess these correlations. Highly correlated features, i.e., those with a correlation coefficient greater than 0.9, were identified. Such features often carry redundant information, and thus, to streamline our data and prevent multicollinearity issues in the subsequent modeling process, these features were removed.

ii. Feature Normalization:
Post correlation analysis, the dataset was normalized using the Standard Scaler. This scaling transformed our features to have a mean of 0 and a standard deviation of 1, ensuring that all features contributed equally to the upcoming processes and models, regardless of their original scale.

iii. Principal Component Analysis (PCA)
PCA was employed as a dimensionality reduction technique. We retained components that accounted for 95% of the variance in the data. This approach allowed us to represent our data in a reduced space while preserving most of its variance, thus balancing the trade-off between data representation and dimensionality.

iv.Random Forest-based Feature Importance:
After PCA, a Random Forest classifier was trained on the normalized dataset. The objective was not for classification per se, but to harness the model's ability to rank features based on their importance in predicting the target variable. A bar plot was then generated, ranking all features based on their importance scores. To further optimize the feature set, only the top 40 features were retained for the subsequent phases of our study.

Feature reduction was instrumental in streamlining our dataset, ensuring efficient and effective modeling in our accented speech recognition tasks for the Malayalam language. This consolidated approach ensured that only the most pertinent features, free of redundancies, were used, laying a solid foundation for the next stages of our research.

## 5 The Exploration of Clustering Techniques for Accented Malayalam Emotional Data

The data illustrates the silhouette scores resulting from applying various clustering algorithms on the scaled dataset of accented speech recognition for the Malayalam language. The silhouette score ranges between -1 and 1, where a high value indicates that the object is well matched to its cluster and poorly matched to neighboring clusters.

i. OPTICS (Ordering Points to Identify the Clustering Structure)
OPTICS is designed to find clusters of various shapes and densities in large datasets without needing to specify the number of clusters beforehand. With a score of 0.4029, it suggests that the clusters are relatively distinct and well-formed. It's an extension of DBSCAN that can find clusters of varying densities. It doesn't require the number of clusters to be specified beforehand.

iii. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)
BIRCH uses a tree structure to hierarchically cluster datasets. A score of 0.1169 suggests that the clusters might be overlapping and not as well-separated.

iii. Ensemble Clustering (Majority Voting)
This approach combines multiple clustering solutions, with a majority voting technique deciding the final cluster. A score of 0.2399 indicates moderate cluster distinction. It combines multiple clustering solutions, and the majority voting technique determines the final cluster.

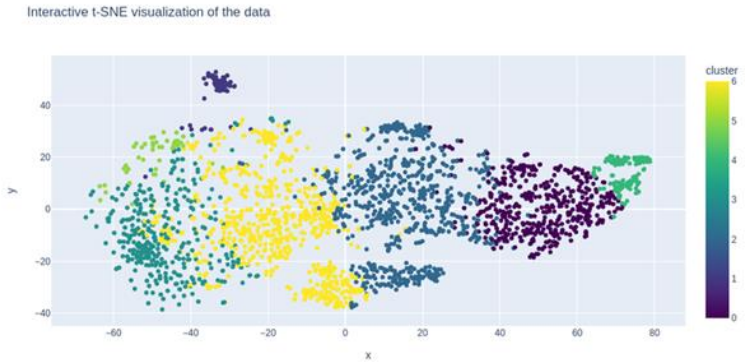


**Fig.5.** The Emotional Clusters

iv.Consensus Clustering
Consensus Clustering aggregates multiple clustering results to find a unified solution. The relatively low score of 0.1191 might indicate overlapping clusters. Builds upon the idea of ensemble clustering but instead of majority voting, it looks for a consensus or common agreement among various clustering solutions. A co-association matrix is created, which is then used for hierarchical clustering.

v. Affinity Propagation
Affinity Propagation identifies clusters through a process of passing messages between pairs of samples until convergence. The highest score obtained in this research is 0.5255, indicates well-defined and distinct clusters. This algorithm doesn't require the predefined specification of the number of clusters in advance.

vi.Mean Shift Clustering
Mean Shift is designed to identify blobs within a smooth density of samples. The obtained score of 0.2511 in this study indicates the presence of distinct clusters. This algorithm operates based on a sliding-window approach, aiming to discover clusters in a smooth density of samples. It utilizes a centroid-based mechanism, updating candidates for centroids to be the means of the data points within a given region.

vii. Agglomerative Clustering
This method represents a form of hierarchical clustering that constructs nested clusters and obtained a score of 0.1053 indicating the possibility of overlapping among the clusters. This mechanism operates by successively merging or splitting clusters in a hierarchical manner.

viii. GMM (Gaussian Mixture Model)
The Gaussian Mixture Model (GMM) operates under the assumption that data points are generated from a mixture of Gaussian distributions. This study obtained score of 0.1368 which indicates the overlapping of clusters and lacking efficiency in clustering the emotional data. GMM is a probabilistic model that posits all data points are generated from a blend of various Gaussian distributions with unknown parameters.

The provided silhouette scores serve as a primary metric to gauge the efficiency of each clustering algorithm. Affinity Propagation, with the highest score, seems to be the most effective for this dataset. Conversely, Agglomerative Clustering has the lowest score, indicating it might not be suitable in this context of study.

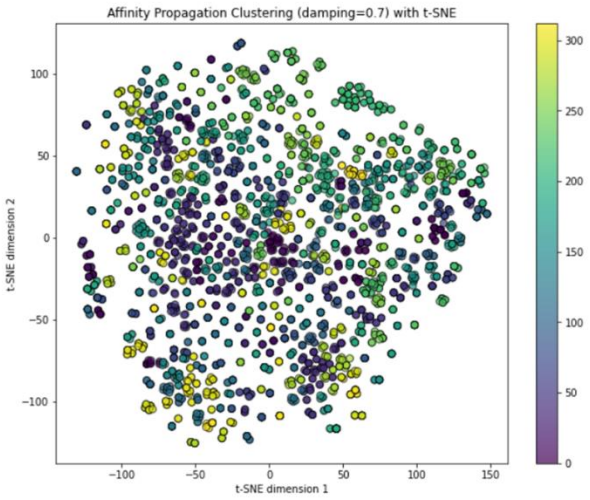


**Fig. 5.** Clusters formed using Affinity Clustering

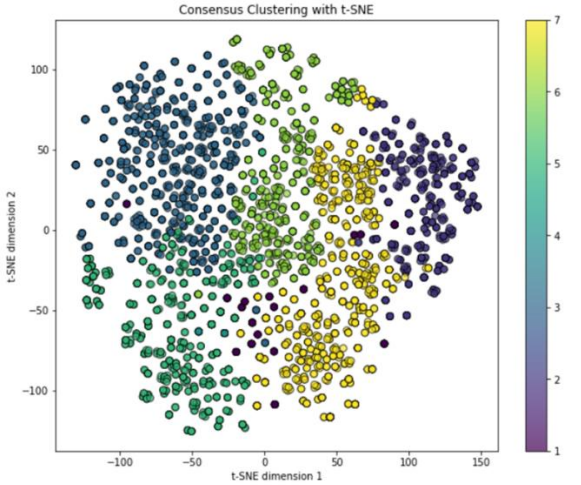


**Fig. 6.** Clusters formed using Consensus Clustering

## 6 Result and Discussion

The clustering algorithms were applied to the dataset of accented speech recognition for the Malayalam language, with the effectiveness of each algorithm being gauged using the Silhouette Score. The summary of the results are:

1. Affinity Propagation yielded the highest silhouette score of 0.5255, suggesting a superior cluster quality with well-defined and distinct groups.
2. OPTICS and Mean Shift Clustering followed suit with scores of 0.4029 and 0.2511, respectively, indicating relatively distinct and well-formed clusters.
3. Ensemble Clustering (Majority Voting) achieved a score of 0.2399, representing moderate cluster distinction.
4. GMM, BIRCH, Consensus Clustering, and Agglomerative Clustering produced silhouette scores below 0.25, suggesting potential overlaps among clusters or less distinct cluster formations.

In the domain of emotional data analysis, the task of clustering emotional expressions is pivotal for gaining insights into the intricate landscape of human sentiment. As the spectrum of emotions is vast and dynamic, the choice of clustering algorithms holds the key to unveiling the hidden structures within this data. This discussion section examines the outcomes of our comprehensive investigation, shedding light on the performance of diverse clustering methodologies. Each algorithm embarked on the challenge of categorizing emotional data, leading to intriguing discoveries and nuanced insights. Within this discourse, we navigate through the standout findings, unravelling the strengths and limitations of each clustering approach, which in turn guide us toward a deeper understanding of emotional data's inherent complexity.

i. Superior Performance of Affinity Propagation:
The remarkable performance of Affinity Propagation in our analysis merits attention. The high silhouette score it achieved stands as a testament to its exceptional ability to identify clusters within the dataset effectively. This achievement can be attributed to the algorithm's innate talent for autonomously determining exemplars in the data. One of the most remarkable aspects is its capacity to negate the need for a predefined number of clusters, allowing it to adapt organically to the intrinsic structure of the emotional data. The superior performance of Affinity Propagation underscores its potential as a robust choice for clustering emotional data, offering a valuable tool for researchers and analysts seeking to explore the nuanced world of human emotions.

ii. Challenges with Hierarchical Clustering:
While Affinity Propagation excelled, the landscape of clustering methodologies was not without its challenges. Both Agglomerative Clustering and BIRCH, representatives of hierarchical methodologies, delivered scores that placed them on the lower end of the performance spectrum. This outcome hints at a potential incongruity between the hierarchical approach and the nature or distribution of the emotional data. It beckons further investigation into the compatibility of hierarchical strategies with the unique characteristics of emotional expressions, providing valuable insights into the complexities of clustering in this context.

iii. Role of Data Distribution with OPTICS:
OPTICS, in contrast, displayed commendable performance, signifying a distinct attribute of the emotional data. Its ability to identify clusters effectively aligns with the presence of clusters of varying densities within the dataset. This insight highlights OPTICS' strength in adaptive clustering, allowing it to detect clusters without the prerequisite of specifying the number of clusters. It provides an essential tool for scenarios where emotional expressions vary widely in density and distribution, opening doors to a deeper understanding of emotional landscapes.

iv.Potential Overlaps with Lower Scores:
The landscape of clustering outcomes presents a challenge embodied by algorithms such as GMM and Consensus Clustering, which yielded lower scores. These results raise questions about potential overlaps within the emotional data or the presence of less distinct cluster formations. To address this challenge, it becomes imperative to embark on further exploration, considering the incorporation of additional evaluation metrics or visualizations. This endeavour promises to shed light on the intricacies of emotional expressions that may not be immediately evident in the clustering results.

v. Ensemble and Consensus Approaches:
Lastly, the Ensemble Majority Voting and Consensus Clustering methods, designed to merge and find common ground among various clustering outputs, offered mixed results. While ensemble techniques often enhance the robustness of clustering results, the specific ensemble or consensus methods employed here exhibited variability in their performance. This underscores the need for fine-tuning and additional parameter optimization to harness the full potential of ensemble methods for emotional data clustering. The mixed outcomes also prompt reflections on the intricacies of combining clustering results and the nuances of consensus-building within emotional data analysis.

**Table 1** SER in Research

| Ref. No | Year | Methodology | Datasets | Key Outcomes |
|---|---|---|---|---|
| 9 | 2023 | Intelligent feature selection with GWO-KNN | Arabic Emirati-accented speech database, RAVDESS, SAVEE | Outperformed traditional methods for recognizing emotions in speech using Grey Wolf Optimizer and K-nearest neighbor classifier. |
| 10 | 2023 | Hybrid Features and CNN-Based Emotion Recognition | Emo-DB, SAVEE, and RAVDESS | A CNN model enhanced with MFCCT features achieved superior performance with accuracy rates of 97%, 93%, and 92% for respective datasets. |
| 11 | 2022 | Emotion Recognition in Urdu with K-nearest neighbors | Corpus collected from 20 subjects | Achieved 66.5% accuracy with K-nearest neighbors. Eliminating the "disgust" emotion led to a 76.5% accuracy improvement. |
| 12 | 2019 | CNN-Based Speech Emotion Recognition System | - | Attained an impressive accuracy of 83.61% on the test dataset, surpassing other methodologies. |
| 13 | 2020 | Comparative Study of SER Systems with Various Classifiers | Berlin and Spanish databases | Different classifiers yielded accuracies ranging from 83% to 94%, and feature selection techniques enhanced performance. |
| 14 | 2020 | Time-Frequency Features using Fractional Fourier Transform | Berlin EMO-DB, SAVEE, PDREC | Proposed method proficiently identified emotional classes with high accuracy across three datasets. |
| 15 | 2020 | Ensemble CNN Model with Multi-Channel EEG and Physiology | DEAP dataset | Enhanced accuracy and stability in emotion recognition using multi-channel EEG and peripheral physiological signals. |
| 16 | 2020 | Two-Stage Approach with Audio Features and Auto-encoder | - | Produced better results in comparison to other SER approaches. |
| 17 | 2020 | Dual-Level Model for SER with MFCC and Mel-Spectrograms | IEMOCAP dataset | Significantly outperformed baseline models and achieved comparable results with multimodal models. |
| 18 | 2019 | BLSTM and Attention Model for SER | - | Outperformed other approaches in terms of accuracy by utilizing speech segments with minimum duration and silence removal threshold. |
| 19 | 2019 | CNN LSTM Networks for Emotion-Related Features | Berlin EmoDB, IEMOCAP | Demonstrated excellent performance in recognizing speech emotion, surpassing traditional methods. |

## 7 Conclusion

In this study, a comprehensive exploration of various clustering algorithms was undertaken to analyse the dataset related to accented speech recognition in the Malayalam language. The Silhouette Score served as the primary metric for evaluating the effectiveness of each clustering methodology.

Notably, Affinity Propagation emerged as the most proficient algorithm, demonstrating its aptitude in discerning distinct clusters within the dataset. On the contrary, some hierarchical clustering methods, despite their widespread applicability, didn't fare as well for this specific dataset, emphasizing the need for tailored algorithmic approaches based on data characteristics. The varied performance across algorithms underscores the significance of understanding underlying data distributions and the inherent characteristics of each clustering method. Furthermore, the results accentuate the pivotal role of ensemble and consensus techniques in potentially enhancing the robustness of clustering outcomes, though their efficacy remains contingent on specific

implementation details. Beyond the numerical evaluations, the intrinsic value of these clusters lies in their potential to provide deeper insights into the nature of accented Malayalam speech.

While this study offers a foundational understanding of the clustering landscape for the dataset at hand, the journey of discovery continues. The road ahead beckons more in-depth analyses, integrative evaluations, and the melding of domain expertise with data-driven insights to revolutionize accented speech recognition for Malayalam.

## References


1. S. Kanwal and S. Asghar, "Speech Emotion Recognition Using Clustering Based GA-Optimized Feature Set," in IEEE Access, vol. 9, pp. 125830-125842, 2021, doi: 10.1109/ACCESS.2021.3111659.
2. Hajarolasvadi N, Demirel H. 3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms. *Entropy*. 2019; 21(5):479. https://doi.org/10.3390/e21050479.
3. A. Dutt and P. Gader, "WaVELEt Multiresolution analysis based Speech Emotion Recognition System using 1D CNN LSTM networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2043–2054, Jan. 2023, doi: 10.1109/taslp.2023.3277291.
4. S. N. Zisad, M. M. Hossain, and K. Andersson, "Speech emotion recognition in neurological disorders using convolutional neural network," in Lecture Notes in Computer Science, Springer Science+Business Media, 2020, pp. 287–296. doi: 10.1007/978-3-030-59277-6_26.
5. A. A. Abdelhamid et al., "Robust Speech Emotion Recognition using CNN+LSTM based on stochastic Fractal Search Optimization Algorithm," IEEE Access, vol. 10, pp. 49265–49284, Jan. 2022, doi: 10.1109/access.2022.3172954.
6. K. S. Lee and H. J. Kim, Design of a convolutional neural network for speech emotion recognition. 2020. doi: 10.1109/ictc49870.2020.9289227.
7. H. A. Abdulmohsin, H. B. A. Wahab, and A. M. J. A. Hossen, "A new proposed statistical feature extraction method in speech emotion recognition," Computers & Electrical Engineering, vol. 93, p. 107172, Jul. 2021, doi: 10.1016/j.compeleceng.2021.107172
8. S. Sunny, D. P. S, and K. P. Jacob, "Discrete wavelet transforms and artificial neural networks for recognition of isolated spoken words," International Journal of Computer Applications, vol. 38, no. 9, pp. 9–13, Jan. 2012, doi: 10.5120/4634-6871.
9. I. Shahin, O. A. Alomari, A. B. Nassif, I. Afyouni, I. A. Hashem, and A. Elnagar, "An efficient feature selection method for arabic and english speech emotion recognition using Grey Wolf Optimizer," Applied Acoustics, vol. 205, p. 109279, Mar. 2023, doi: 10.1016/j.apacoust.2023.109279.
10. A. S. D. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati, "Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network," Applied Sciences, vol. 13, no. 8, p. 4750, Apr. 2023, doi: 10.3390/app13084750.
11. A. Asghar, S. Sohaib, S. Iftikhar, M. Shafi, and K. Fatima, "An Urdu speech corpus for emotion recognition," PeerJ, vol. 8, p. e954, May 2022, doi: 10.7717/peerj-cs.954.
12. A. B. A. Qayyum, A. Arefeen, and C. Shahnaz, Convolutional Neural Network (CNN) Based Speech-Emotion Recognition. 2019. doi: 10.1109/spicscon48833.2019.9065172
13. L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, and C. Cléder, "Automatic Speech Emotion Recognition using Machine learning," in IntechOpen eBooks, 2020. doi: 10.5772/intechopen.84856.
14. S. Langari, H. Marvi, and M. Zahedi, "Efficient speech emotion recognition using modified feature extraction," Informatics in Medicine Unlocked, vol. 20, p. 100424, Jan. 2020, doi: 10.1016/j.imu.2020.100424.
15. H. Huang, Z. Hu, W. Wang, and M. Wu, "Multimodal Emotion Recognition Based on Ensemble Convolutional Neural Network," IEEE Access, vol. 8, pp. 3265–3271, Jan. 2020, doi: 10.1109/access.2019.2962085.
16. H. Aouani and Y. B. Ayed, "Speech Emotion Recognition with deep learning," Procedia Computer Science, vol. 176, pp. 251–260, Jan. 2020, doi: 10.1016/j.procs.2020.08.027.
17. J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, Speech Emotion Recognition with Dual-Sequence LSTM Architecture. 2020. doi: 10.1109/icassp40776.2020.9054629.
18. B. T. Atmaja and M. Akagi, Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model. 2019. doi: 10.1109/icsigsys.2019.8811080.
19. J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," Biomedical Signal Processing and Control, vol. 47, pp. 312–323, Jan. 2019, doi: 10.1016/j.bspc.2018.08.035.