

A Multimodal Speech Emotion Recognition Framework for Malayalam through Audio-Text Fusion

Athira Raj^{1,*}, Christy James Jose², and Biju K. S.²

¹Department of Computer Science and Engineering, College of Engineering Trivandrum, Thiruvananthapuram, India

²Department of Electronics and Communication Engineering, College of Engineering Trivandrum, Thiruvananthapuram, India

*Corresponding author: athiraraj673@gmail.com

Abstract. Speech emotion recognition (SER) is used in many domains, such as translation, intelligent assistants, healthcare monitoring, large language models, and human-computer interaction. Emotion recognition in Malayalam, however, remains challenging because of the language's rich morphological structure. This work introduces a multimodal SER model for Malayalam that uses both speech and text information. Because few Malayalam emotion databases are available, a dataset was constructed from publicly available audiovisual content. The method uses a wav2vec 2.0 model to extract audio features, while the text features are obtained with a pretrained language model. The unified model then fuses the audio and text features for emotion prediction. Label consistency in the constructed dataset was further examined with an embedding-based analysis. On a dataset of 1,040 samples evenly distributed across four emotions, the proposed model achieved an accuracy of 78.85% and a macro-averaged F1-score of 0.79 on the test set, outperforming an audio-only baseline.

Keywords. Audio-text fusion; Malayalam language; multimodal learning; speech emotion recognition; wav2vec 2.0

I. INTRODUCTION

Speech emotion recognition (SER) has become an integral part of human-computer interaction, and many SER methods based on machine learning and deep learning have been proposed over the past decade. Traditional SER approaches relied on handcrafted acoustic features and conventional machine learning algorithms [1], which required extensive feature engineering and generalized poorly. Deep learning has advanced SER considerably by enabling models such as convolutional and recurrent neural networks to learn speech features automatically, which has improved performance. Multimodal emotion recognition has gained importance because emotions are expressed not through speech alone but also through semantic and contextual information. Combining speech with textual features therefore provides richer emotional cues than conventional single-modal approaches [2].

A multimodal emotion recognition framework that uses large language models for emotion classification is proposed in [3]. It uses a rule-based system to convert nonverbal cues, namely audio features such as loudness and visual features such as facial expressions, into a unified textual format. Self-supervised models such as wav2vec 2.0, HuBERT, and WavLM have been used to improve automatic speech recognition (ASR) for low-resource languages that lack labeled datasets [4]. A unified scheme for multilingual speech emotion conversion that transforms neutral speech into expressive tones is presented in [5]. That study maps vocal features across German, Telugu, and English

with a mixed-lingual training approach that does not require parallel training data, and it shows that emotional characteristics can be transferred across language boundaries even in low-resource scenarios.

The feasibility of multilingual and cross-lingual SER was evaluated on English and French datasets with an attentive convolutional neural network (CNN) model [6]. The authors showed that training on several languages at once is viable and that arousal (intensity) transfers across languages better than valence (positivity or negativity). An SER method that addresses the scarcity of labeled data in low-resource languages through cross-domain adaptation is presented in [7]. Its architecture consists of a speech recognition front end built on pretrained models and a specialized back-end network that maps linguistic features into the emotional domain, and the approach was demonstrated on German and Thai datasets.

The Kazakh Emotion Multimodal Optimizer (KEMO) is a framework for SER in the Kazakh language that merges text-based semantic analysis of translated datasets with audio-based acoustic features processed by deep learning architectures [8]. To overcome the challenges of a low-resource language, that work also presents a dynamic weighting method that prioritizes either the vocal features or the meaning of the text, depending on the clarity of the input signal.

Malayalam is a Dravidian language native to the Indian state of Kerala. Its rich morphological structure affects the way emotions are expressed in speech, and research on SER

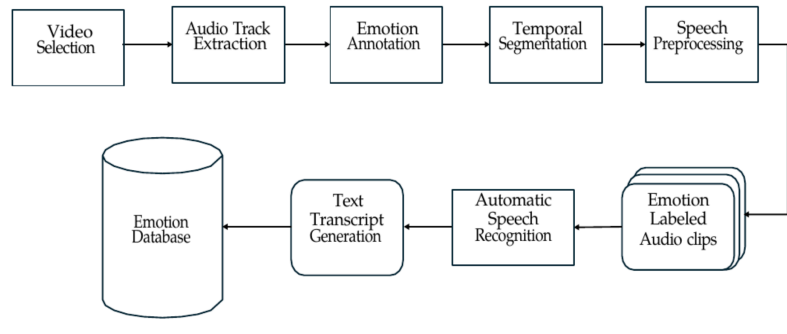


Fig. 1. Procedure for constructing the emotion-labeled speech dataset.

for Malayalam remains limited. A hybrid Gaussian mixture model and deep belief network (GMM-DBN) architecture for Malayalam SER is presented in [9]. The authors compared recordings of emotionally biased text and neutral sentences to study their effect on the recognition of emotions such as anger, happiness, and sadness. Using Mel frequency cepstral coefficients (MFCCs) processed by DBNs, the study showed that emotionally biased prompts substantially improve recognition. However, that work relies only on acoustic features, without linguistic information, and therefore lacks cross-accent generalization; its evaluation was also performed in a speaker-dependent setting.

This work presents a deep learning model that predicts emotions from Malayalam speech by combining audio and textual information. Speech embeddings capture prosodic cues related to emotion, while textual embeddings contribute contextual information from the transcripts, and the two modalities are merged by feature-level fusion for emotion classification. The dataset was constructed from speech audio extracted from publicly available sources, manually labeled with emotion categories, and used for model training. In addition, an embedding-based visualization was used to analyze the consistency of the manually assigned emotion labels.

The remainder of this paper is organized as follows. Section II presents the proposed method for SER in Malayalam. Section III describes the experimental setup and reports the results. Section IV concludes the paper.

II. PROPOSED METHOD

This section explains the proposed multimodal approach to SER in Malayalam. The method integrates acoustic and textual information and uses an embedding-based analysis to validate the manually annotated dataset.

A. Dataset

The dataset was constructed from selected Malayalam movies. Fig. 1 shows the procedure for constructing the speech emotion database. Emotion labels were assigned manually on the basis of contextual and vocal cues. The corresponding speech clips, with an average duration of approximately 6 s, were extracted from the movie audio. Several preprocessing steps were applied to improve the quality of the collected speech samples: the speech signal was processed to isolate the speaker's voice, and noise filtering was performed to suppress background noise. All audio samples were resampled to 16 kHz before feature extraction. A textual transcript was generated for each speech utterance with

an ASR algorithm. A total of 1,040 labeled samples were obtained across four emotion classes (happy, sad, angry, and neutral), with 260 samples per class.

B. Feature Extraction

Emotion-relevant features were extracted with separate models for the speech signal and the textual transcript. The audio waveform $x(t)$ was processed with the pretrained wav2vec 2.0 model [10] to obtain vector embeddings:

$$Z = f_{\text{wav2vec}}(x(t)) = \{z_1, z_2, \dots, z_T\}, \quad z_i \in \mathbb{R}^{768}, \quad (1)$$

where T is the number of frames in the utterance. Each embedding z_i is a vector that captures spectral patterns, prosodic cues, and contextual information from the frames through the transformer layers. The embedding dimension of 768 corresponds to the hidden representation size of the wav2vec 2.0 transformer encoder. To obtain a fixed-length representation of the whole utterance, the frame-level embeddings were combined by max pooling:

$$Z_{\text{utterance}} = \max_{i=1, \dots, T} z_i. \quad (2)$$

The text of each speech sample was obtained as a transcript from the speech recognition model. The emotional features and contextual information in the transcript were then extracted with MuRIL (Multilingual Representations for Indian Languages) [11], a multilingual transformer trained with masked language modeling. Each token w_i , the smallest meaningful segment of text, is mapped to a contextual embedding:

$$H = f_{\text{MuRIL}}(w_1, \dots, w_N) = \{h_1, \dots, h_N\}, \quad (3)$$

where each $h_i \in \mathbb{R}^{768}$ represents the semantic meaning of the word, its relationship to nearby tokens, and the emotion information present in the sentence. The final textual representation is the embedding of the [CLS] token at the output of the transformer:

$$H_{\text{sentence}} = h_{[\text{CLS}]}. \quad (4)$$

The speech and textual representations are then concatenated to form the multimodal feature vector

$$F = [Z_{\text{utterance}}; H_{\text{sentence}}]. \quad (5)$$

C. Model Architecture

A feedforward neural network with two fully connected hidden layers of 256 and 128 units was used to predict emotional states from the multimodal features obtained from the speech and text embeddings. Each hidden layer is followed by a rectified linear unit (ReLU) activation and a dropout layer with a rate of 0.3. The multimodal feature vector F is

the input to the classification network, which maps it through a linear output layer to a four-dimensional output corresponding to the labeled emotion categories. The model was trained with the cross-entropy loss. Fig. 2 shows the proposed architecture, and Algorithm 1 summarizes the procedure.

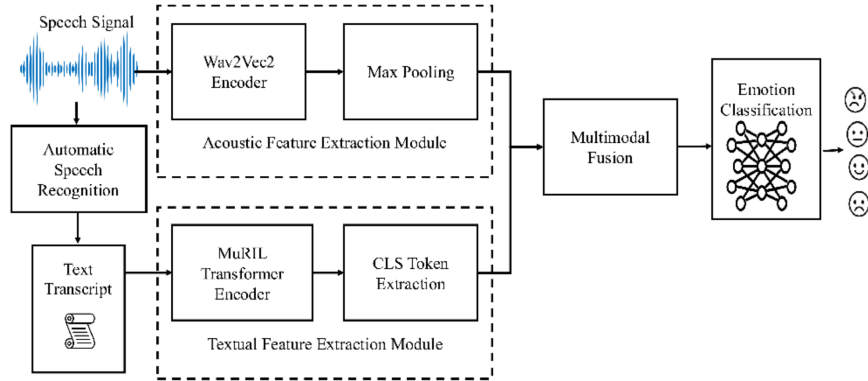


Fig. 2. Architecture of the proposed speech emotion recognition framework.

III. RESULTS AND DISCUSSION

Malayalam remains a low-resource language with limited annotated datasets and linguistic tools, and its agglutinative morphology and phonetic richness further increase the complexity of SER. Experiments were conducted on the manually annotated Malayalam speech emotion dataset of 1,040 utterances across four emotion categories: happy, sad, angry, and neutral. The dataset was divided into training, validation, and test sets with the class distribution preserved. The model was trained with the Adam optimizer at a learning rate of 0.0001 and a batch size of 16 for 30 epochs. The effectiveness of the proposed multimodal framework was evaluated by comparison with an audio-only baseline, and test-set performance was measured by overall accuracy and macro-averaged F1-score. Table 1 compares the overall performance of the two models.

TABLE 1. COMPARISON OF AUDIO-ONLY AND MULTIMODAL MODEL PERFORMANCE

Model	Accuracy (%)	Macro F1-score
Audio-only	62.98	0.63
Multimodal	78.85	0.79

Table 2 lists the accuracy reported by existing SER methods on their respective datasets. Because the datasets differ, the figures are not directly comparable with one another or with the present results.

TABLE 2. PERFORMANCE OF RELATED SER METHODS

Method	Dataset	Accuracy (%)
DBMER [1]	IEMOCAP	74.80
KEMO [8]	ENU KEMO	77.10
GMM-DBN [9]	Malayalam biased speech data	77.29
Proposed model	Malayalam movie speech dataset	78.85

A class-wise analysis was performed to examine the effect of multimodal fusion on individual emotion categories.

ALGORITHM 1. PROPOSED MULTIMODAL EMOTION RECOGNITION FRAMEWORK

```

1: for each audio-transcript pair  $(x, W)$  do
2:   Extract the speech embedding with wav2vec 2.0
3:   Extract the text embedding with MuRIL
4:   Concatenate the embeddings (feature-level fusion) to form  $F$ 
5:   Classify  $F$  into one of the emotion categories
6: end for

```

Table 3 compares the F1-scores of the audio-only baseline and the proposed multimodal framework. The class-wise results show a consistent improvement across all emotion categories when audio and text features are fused. The largest gains were observed for sad and happy, the two classes with the lowest scores in the audio-only baseline. Fig. 3 shows the training and validation accuracy of the proposed model over the epochs.

TABLE 3. CLASS-WISE F1-SCORE COMPARISON

Emotion	Audio-only	Multimodal
Sad	0.59	0.80
Happy	0.56	0.75
Angry	0.64	0.77
Neutral	0.74	0.83
Macro average	0.63	0.79

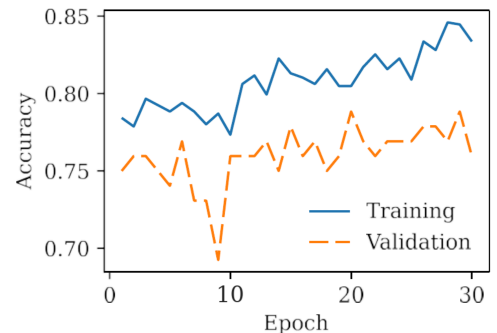


Fig. 3. Training and validation accuracy versus epoch.

Table 4 gives the confusion matrix of the proposed multimodal model on the test set.

A qualitative analysis of the learned embedding space was performed with dimensionality reduction. Fig. 4 shows a uniform manifold approximation and projection (UMAP) visualization of the audio embeddings [12], which was used to analyze how the manually annotated emotion categories are distributed in the embedding space. The visualization

shows partial clustering of the emotion categories, with overlap between closely related emotions.

TABLE 4. CONFUSION MATRIX OF THE PROPOSED MULTIMODAL MODEL ON THE TEST SET

Actual / Predicted	Sad	Happy	Angry	Neutral
Sad	45	1	5	1
Happy	5	41	3	3
Angry	5	9	38	0
Neutral	5	6	1	40

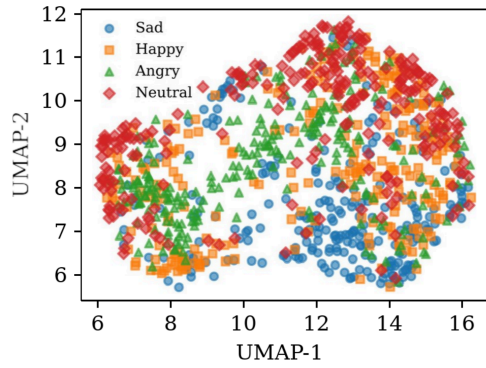


Fig. 4. UMAP visualization of the audio embeddings for the four emotion classes (happy, sad, angry, and neutral).

IV. CONCLUSION

This work presented a multimodal SER model for Malayalam that integrates audio and textual information. The experimental evaluation showed that multimodal integration improves classification performance and gives more consistent emotion recognition than an audio-only model. Despite these improvements, overall performance is limited by the dataset size and by the set of emotion categories considered. Expanding the dataset with additional emotion categories and more varied speech samples may further improve performance. The present study is based on a manually annotated Malayalam movie speech dataset acquired specifically for this work.

As an extension, future research may explore semi-automated labeling and validation methods, such as third-party audio verification mechanisms inspired by CAPTCHA-style tasks, to improve the reliability of annotations during dataset development. Such cross-annotation and validation by multiple human annotators, together with a sufficiently large dataset, would make emotion annotation more consistent, especially for speech models in low-resource languages.

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and preparation of this work and approved the final version of the manuscript.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

DECLARATION OF GENERATIVE AI USE

The authors declare that no generative artificial intelligence tools were used in the preparation of this manuscript. The authors take full responsibility for the content of the publication.

DATA AVAILABILITY

The data supporting the findings of this study are available from the corresponding author upon reasonable request. The audio clips were drawn from commercial Malayalam films.

ETHICS STATEMENT

This study used publicly released film audio and did not involve direct interaction with human participants or animals; ethical approval was therefore not required.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- [1] Samuel Kakuba, Alwin Poullose, and Dong Seog Han, "Deep learning approaches for bimodal speech emotion recognition: Advancements, challenges, and a multi-learning model," *IEEE Access*, vol. 11, pp. 113769-113789, 2023, doi: 10.1109/ACCESS.2023.3325037.
- [2] Sepideh Kalateh, Luis A. Estrada-Jimenez, Sanaz Nikghadam-Hojjati, and Jose Barata, "A systematic review on multimodal emotion recognition: Building blocks, current state, applications, and challenges," *IEEE Access*, vol. 12, pp. 103976-104019, 2024, doi: 10.1109/ACCESS.2024.3430850.
- [3] Omkumar Chandraumakantham, N. Gowtham, Mohammed Zakariah, and Abdulaziz Almazayad, "Multimodal emotion recognition using feature fusion: An LLM-based approach," *IEEE Access*, vol. 12, pp. 108052-108071, 2024, doi: 10.1109/ACCESS.2024.3425953.
- [4] Jing Zhao and Wei-Qiang Zhang, "Improving automatic speech recognition performance for low-resource languages with self-supervised models," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1227-1241, 2022, doi: 10.1109/JSTSP.2022.3184480.
- [5] Susmitha Vekkot and Deepa Gupta, "Fusion of spectral and prosody modelling for multilingual speech emotion conversion," *Knowl.-Based Syst.*, vol. 242, Art. no. 108360, 2022, doi: 10.1016/j.knosys.2022.108360.
- [6] Michael Neumann and Ngoc Thang Vu, "Cross-lingual and multilingual speech emotion recognition on English and French," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, 2018, pp. 5769-5773, doi: 10.1109/ICASSP.2018.8462162.
- [7] Konlakorn Wongpatikaseree, Sattaya Singkul, Narit Hnoohom, and Sumeth Yuenyong, "Real-time end-to-end speech emotion recognition with cross-domain adaptation," *Big Data Cogn. Comput.*, vol. 6, no. 3, Art. no. 79, 2022, doi: 10.3390/bdcc6030079.
- [8] Mamyr Altaibek, Altanbek Zulkhazhaz, Banu Yergesh, Gulmira Bekmanova, and Tileukhan Aibol, "A multimodal framework for speech emotion recognition in low-resource languages," *J. Artif. Intell. Technol.*, vol. 5, pp. 354-364, 2025, doi: 10.37965/jait.2025.0781.
- [9] Athira Chandran, D. Pravena, and D. Govind, "Development of speech emotion recognition system using deep belief networks in Malayalam language," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Udupi, India, 2017, pp. 676-680, doi: 10.1109/ICACCI.2017.8125919.
- [10] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 12449-12460. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html>
- [11] Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar, "MuRIL: Multilingual representations for Indian languages," arXiv:2103.10730, 2021, doi: 10.48550/arXiv.2103.10730.
- [12] Leland McInnes, John Healy, and James Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," arXiv:1802.03426, 2018, doi: 10.48550/arXiv.1802.03426.