INTERNATIONAL JOURNAL OF APPLIED ARTIFICIAL INTELLIGENCE AND ROBOTICS
E-ISSN: 3124-1212, DOI: 00.000/XXXX.YYYY.II.NNN
Vol. 2 No. 1 2026 pp. 01-10

Evaluating the Impact of Behavioural Features on Hindi Speech Emotion Recognition: A Multimodal Deep Learning Approach

Sujata Kotian[1], Santosh Singh[2*]

[1,2] Department of Information Technology, University of Mumbai

*Corresponding Author: sksingh@tcsc.edu.in

Article info: Received 07-01-2026, Revised 30-01-2026, Accepted 15-02-2026

Abstract

Context: Speech Emotion Recognition (SER) is an important part of affective computing, but that it cannot work effectively in low resource languages like Hindi. The available SER systems have focused on low-level speech features (acoustic and prosodic) and little has been done to investigate the high-level behavioral speech features (e.g., pauses and rhythm) even though they are significant in human emotional communication.
Objective: This study aimed to explore the hypothesis, whether explicit behavioral speech features can enhance Hindi SER performance, as well as study their joint role in complementing acoustic and prosodic features within a multimodal deep learning system.
Method: A curated Hindi emotional speech corpus of 2,370 utterances of 25 speakers composed of seven emotion classes was studied through a controlled experimental study. The acoustic, prosodic and behavioral features were obtained and represented with a dual-branch multimodal deep learning framework that included CNN/transformer and BiLSTM-attention modules.
Results: The entire multimodal model had an accuracy of 83.9% and a macro-F1 of 0.81, which was significantly higher than the acoustics-only and acoustics-prosodic baselines. The behavioral features provided significant progress to low-arousal emotions, including sadness and neutral, and medium to large effect sizes.
Conclusions: The results show that Hindi SER accuracy and strength is significantly increased by behavioral speech cues. To practitioners, the findings provide justification to apply behavior-aware SER in practice, whereas to researchers, they show the necessity to explicitly model the behavioral characteristics in low-resource and culturally diverse languages.

Keywords: Hindi Speech; Emotion Recognition, Behavioral Speech Features, Multimodal Deep Learning, Prosody; Acoustic Modelling

1. Introduction

Emotion is also one of the most basic dimensions of human communication in that it defines how individuals convey intentions, understand social messages and establish effective interactions. With the advent of intelligent systems in our day to day activities, the capability of the machine to detect and react accordingly to human emotions has become one of the focal issues in affective computing. The initial developments in the discipline were based on single-modal cues like facial expressions or speech, but more complex, context-sensitive models that can decode emotional states in a wide variety of real-world contexts have continued to emerge. This transition has been facilitated by the great amount of advancement in machine learning and signal processing that has enabled researchers to investigate how a combination of speech, language, and behavior patterns forms the patterns of emotional communication.

Speech Emotion Recognition (SER) has been developing in this space as an important research direction because of the centrality of speech in human interaction. The main notion of SER is to deduce latent emotional conditions with the help of changing acoustic, prosodic and articulatory features. Previous research focused mainly on spectral features -including MFCCs, pitch, jitter, shimmer or energy- used together with traditional classifiers. Such surveys as those by Koolagudi and Rao (2012) and the extensive review by Dar and Delhibabu (2024) demonstrate that SER has also developed over time to more complex representation learning models than the simple handcrafted features. With this improvement, speech-based systems continue to be challenged, especially by the speaker heterogeneity, spontaneous emotion, data unavailability, or noise of the environment. The INTERSPEECH Emotion Challenge (Schuller et al., 2009) one of the most popular stress tests in this sphere, was used to show how even the most sophisticated models can be confused with the problem of emotional variability when training and testing conditions are not matched.

These problems are magnified in low-resource languages and non-English languages as emotion corpora are annotated and pattern of emotion expression varies significantly. Hindi is one of the most spoken languages in the world, which has specific phonetic and prosodic peculiarities which affect the articulation of emotions. Although a few works have been conducted on Indian languages, the work at hand is quite limited. As an example, Pawar and Patel (2015) examined the MFCC based recognition of Hindi emotions with the help of DTW, and other regional studies, such as Ibn Nasr et al. (2025) on the dialect of Tunisia show how cultural and linguistic structures can create new behavioral and acoustic patterns. Large, high quality Hindi emotional datasets are limited thus limiting systematic experimentation, especially where multimodal fusion or analysis of behavior is needed. Hindi resources are modest in comparison to mainstream corpora such as IEMOCAP (Busso et al., 2008) and it is not possible to develop resilient generalizable models. Simultaneously, the general emotion recognition field experienced a phase of unquestionable shift towards multimodal and deep learning architecture. The combination of speech, text, and facial behavior cues, as depicted in the review by Poria et al. (2017), has always been found to enhance emotions prediction. More modern innovations show that deep networks can acquire complementary intermodal cues; in one case, Mittal et al. (2020) suggested an M3ER framework, which fuses facial, textual and speech features based on multiplicative fusion to achieve significant performance gains. Transformer based processing is becoming more popular in contemporary models, such as Dhal et al. (2024) and Bhoite (2025) both demonstrate the possibility of using multimodal or speech-driven transformers on constrained or real-time systems. These trends indicate the increased awareness of the fact that emotional communication is always multimodal, and that at human level, recognition needs to be captured at that complexity.

One of the most promising approaches to affective computing is the study of behavioral features Speech derived cues which are expressive patterns in addition to traditional acoustics. Behavioral indicators may involve the rate of speaking, pauses, hesitation, emphasis, turn taking behavior and other time related information which indicates underlying affective or cognitive conditions. Although classical SER studies focused on acoustics, behavioral signal processing has revealed that human speech contains a rich amount of information regarding social interaction, emotional control and even psychological well being. In spite of the fact that there are certain multimodal frameworks that implicitly reflect behavioral patterns, there has been very little systematic assessment of explicit behavioral features particularly with languages such as Hindi. Both the reviews by Dar and Delhibabu (2024) and the conceptual argumentation of the grounding and communication introduced by van der Velde (2015) imply that emotional interpretation is based not only on the content of the signals but also on the way speakers control their behavior in the process of communication.

Although there is an increase in interest in multimodal and behavioral approaches, there still remains a gap in the literature. To begin with, although research using multimodal systems like M3ER (Mittal et al., 2020) and surveys like Poria et al. (2017) prove the usefulness of fusion, there is hardly any research using such principles on the Hindi emotional speech. The available literature on the Hindi studies is based on small datasets, and it is mainly concerned with acoustic modeling without exploring the role of behavioral cues in recognition performance. Second, most SER models have been trained on acted datasets or Western corpora; it is not clear whether they can be applied to natural or culturally based Hindi speech. Third, there are few explicit feature level comparisons, comparing the interaction between behavioral cues and spectral or prosodic features. Bhoite (2025) and Dhal et al. (2024) mention contemporary architectures that can be used in real time, but neither provides a direct analysis of the behavioral information that affects the model robustness. Lastly, the available literature lacks the necessary information on how to build multimodal Hindi SER pipelines in an environment, which simulates real communication scenarios.

The current research fills these knowledge gaps by examining the effect of behavioral speech features during the multimodal deep learning model to Hindi Speech Emotion Recognition. This work is unlike the traditional systems which mostly make use of spectral or prosodic inputs, explicit behavioral descriptors used in it are the pause distribution, variation in rhythm, fluctuation in speaking rate, and emphasis pattern in addition to acoustic features. It is the goal to measure the enhancement of emotion representation by behavioral information, as well as to find out whether it is possible to fill in data constraints commonly experienced in Hindi SER with such cues.

There are four objectives of the study:

- to come up with a multimodal SER pipeline which takes into consideration acoustic, prosodic and behavioral features of Hindi emotional speech;
- to assess the role of behavioral qualities using controlled ablation procedures;
- to make comparisons of unimodal and multimodal configurations in various categories of emotion;
- to evaluate model strength in spontaneous and acted emotional circumstances where possible.

Overall, this article makes a contribution to Hindi SER literature in terms of providing a systematic and behavior-conscious multimodal framework and in developing an empirical evidence of the worth of behavioral speech cues in emotional modeling. The rest of this paper will be structured in the following way: Section 2 will be reviewing the relevant literature on speech based and multimodal emotion recognition. Section 3 reports the dataset, the strategy of feature extraction, and the deep learning structure. Experimental procedures and metrics of evaluation are recorded in section 4. The results are presented and interpreted in Section 5, and the conclusion is made in Section 6, where some implications and future work directions are provided.

2. Related Work

2.1 Foundations of Audio–Visual and Speech-Based Emotion Modelling

The field of computational emotion analysis is growing dramatically now that deep learning tools are made accessible that can detect subtle emotional elements of speech and visual representations. The hybrid deep model suggested by Zhang et al. (2017) [16] can be regarded as one of the influential studies in this direction as it proved that the joint learning of audio-visual representations is more effective as an encoder of affective signals compared to modalities trained independently. The significance of their work was that they could model the timing of temporal synchrony between speech dynamics and facial expressions a concept that other multimodal systems have improved upon in future. Simultaneously with multimodal, early speech based systems had studied the use of convolutional architecture to derive emotion sensitive features on raw spectrograms. The study conducted by Mao et al. (2014) [17] determined that CNNs can be trained on salient spectral patterns of emotional states thus eliminating the need to adopt manually designed descriptors. On the same note, Fayek et al. (2017) [18] evaluated a number of deep architectures and indicated how the model choice, activation functions, and input normalization have significant effects on SER performance. All these studies provided a solid methodological foundation to further work, demonstrating that under the appropriate settings deep networks can be relied upon to substitute handcrafted features.

2.2 Advances in Deep Learning Architectures for Speech Emotion Recognition

As end to end modeling expanded, scholars started applying representation learning as part of the SER pipeline. According to the survey by Prabhavalkar et al. (2023) [19], the development of end to end speech recognition models such as encoder decoder models and CTC based systems was illustrated, and the systems affected the performance of other tasks such as emotion recognition in which continuous speech conditions and spontaneous variations are common. Their results showed that it is crucial to have strong acoustic modeling in the case of utterances that are rich in emotion. In line with this, Latif et al. (2020) [20] gave a thorough explanation of deep representation learning in speech processing, which has weaknesses, including data imbalance, domain mismatch, and speaker variability. They can be directly applied to the low resource setting of emotion recognition, where the expression of emotions differs greatly among speakers and dialects. New SER innovations were the result of the appearance of transformer architectures. Similar to Akinpelu et al., [26] Vision Transformers trained to classify speech by spectrogram embeddings are capable of recognizing fine grain emotional signals and performing better than standard CNN RNN hybrids. In the meantime, Jiang et al. (2021) [25] suggested a multi attention CRNN architecture that can adequately assess the spectral patterns and the evolution of emotions over time. The overall picture of these works is the tendency in the direction of architectures incorporating attention mechanisms to optimize the extraction of emotional features.

2.3 Multimodal Emotion Recognition and Deep Fusion Frameworks

Recognition of emotion has also moved away to multimodal fusion models which are capable of sensing richer affective information. One of the earliest time end to end multimodal deep networks was introduced by Tzirakis et al. (2017) [24] and reported to be notably better in robustness and generalization. More recent approaches, including the MSER framework by Khan et al. (2024) [21] [21], used cross attention to combine multimodal embeddings to enable the model to selectively weight features in each modality. Their findings emphasized the importance of deep fusion strategies in reducing ambiguity in the emotional cues is particularly relevant when dealing with spontaneous or cultural variable speech. Generalized overviews like the survey by Lian et al. (2023) [22] summarized those developments by visualizing how multimodal approaches have been developed and evolved in terms of speech, text, and facial modalities. They also pointed to endemic issues that include modality imbalance, synchronization and cultural variability; of particular importance to the language like Hindi where emotional prosody and behavioral cues could vary very much in comparison to the Western corpora. Regarding the application perspective, Feld et al. (2019) [23] reviewed software platforms to create multimodal interactive systems and noted the necessity of noisy and real world data architecture. Their observations support the need to develop SER models that can be resistant to different linguistic, acoustic, and behavioral conditions.

2.4 Behavioral Cues, Prosody, and Cognitive-Affective Interpretation

In addition to spectral and prosodic attributes, the recent research has emphasized the importance of behavioral cues, including hesitation, pause organization, stress patterns, turn taking behavior and speech rhythm in the meaning of affective states. Handy evidence to support the assertion that, prosodic contours and speech

prosody bear more informative behavioral persuasions, which can shape emotion interpretation, was presented by Sasu (2025) [29] in language where behaviors are culturally influenced by expressive patterns.
Other related insights into complementary views of cognitive science, including Dijkstra and Peeters (2023) [30], also state that emotionally based communication is anchored not only in the acoustic expression, but also in the bodily behavior and anticipated expectation on the side of the listener. This school of thought has been found to be very similar to recent behavioral signal processing methods where speech behavior is considered a result of cognitive load, as well as, emotional state.

These concepts are indirectly supported by deep learning studies. As an example, Zhang et al. (2017) [16] revealed that audio visual data affective cues tend to rely on synchrony in behavior, as opposed to singular spectral cues. Similarly, Mao et al. (2014) [17] and Jiang et al. (2021) [25] showed that through the feature of time modeling, when used with attention, the model inherently reveals nuances in speech behavior that can be found in the rhythms and pauses of speech. In spite of this, explicit modeling of behavioral speech characteristics is still an issue in most SER systems, especially with Hindi and other low resource languages. This gap also stimulates the necessity of the systematic assessment of the role of the behavioral cues in the effect of the emotion recognition performance.

2.5 Emerging Challenges and Identified Gaps in Multimodal and Behavioral SER

Despite the major advances achieved in multimodal and end to end SER, it is possible to outline several loopholes in the work examination. First, the present multimodal systems (e.g., Khan et al., 2024 [21]; Tzirakis et al., 2017 [24]) are oriented on the concept of audio visual fusion, and there is not much consideration of the behavioral speech clues. This forms an essential loophole in comprehending how turn taking and stopping, along with the speed of speech, impact the expression of emotions, particularly because the direct evidence of prosodic behavioral interrelation is high as reported by Sasu (2025) [29]. Second, although transformer based and attention guided CNN–BiLSTM (Akinpelu et al., 2024 [26]) architectures demonstrate better results, they seldom utilize behavioral descriptors as direct inputs. Rather, behavioral cues are usually acquired implicitly, and it is hard to measure their unique input. Third, it has been noted in the current studies (Lian et al., 2023 [22]; Latif et al., 2020 [20]) that low resource languages are characterized by distinctive emotional variability challenges, but few studies address them in multimodal or behavior aware approaches. This is especially so in the case of Hindi where culturally sensitive prosody and expression of emotion are different and not similar to those of commonly used English standards. Lastly, numerous deep learning analyses (Fayek et al., 2017 [18]; Prabhavalkar et al., 2023 [19]) point to the vulnerability of SER models to training conditions, noise, and imbalance in the dataset. Such constraints strengthen the necessity of structures that combine acoustic, prosodic, and behavioral data to enhance the generalization.

3. Method

This part explains the entire methodological pipeline that was used to assess how behavioral speech features can improve Hindi Speech Emotion Recognition (SER) within a multimodal deep learning framework. The overall methodological structure is in line with standard Scopus index conventions and comprises the stages of dataset preparation, preprocessing, feature engineering, multimodal network design, training configuration, and statistical evaluation strategies. The entire process is depicted in Figure 1, while the description of the dataset and the feature taxonomies are presented in Table 1 and Table 2, respectively.

3.1 Research Design and Overall Framework

The investigation in question is based on an experimental, data driven design which is aimed at quantifying the independent and combined effects of acoustic, prosodic, and behavioral speech descriptors. The procedure starts with the procurement of a curated Hindi emotional speech dataset, and then it is followed by the preprocessing steps which include normalization, segmentation, and voice activity detection. Once the data has been preprocessed, three separate feature streams are derived acoustic features (e.g., MFCCs, log Mel spectrograms), prosodic features (e.g., pitch, duration), and behavioral cues (e.g., pause ratio, speech rate, rhythmic variability).

To create short time frames used for acoustic analysis, the speech signal is segmented using:

Equation 1 (Short Time Framing)

$$x_t[n]=x[n]\cdot w[n-tH] \tag{1}$$

These 3 streams are subsequently input into a multimodal deep learning architecture that consists of an acoustic convolutional branch and a prosodic behavioral recurrent branch. The representations from both streams are concatenated with the help of a fusion layer and fed to the final classification head. The evaluation of the performance is done through four ablation settings to measure the impact of behavioral features in isolation. Figure 1, "Overall Proposed Framework for Behavior Aware Hindi Speech Emotion Recognition," briefly depicts this architecture where each block represents one step of the pipeline.

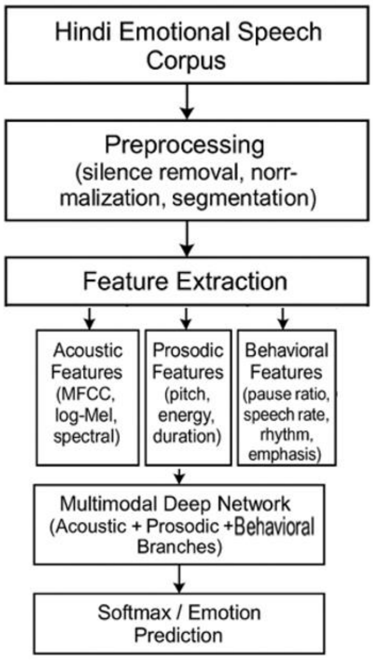

Figure 1. Overall Proposed Framework for Behavior Aware Hindi Speech Emotion Recognition

3.2 Dataset Description and Preprocessing

3.2.1 Hindi Speech Emotion Corpus

The dataset for this research was based on recorded emotional Hindi spoken utterances that cover the major expressive categories typically used in SER studies, i.e., anger, happiness, sadness, fear, disgust, surprise, and neutral speech. The corpus comprises the recordings of the adult speakers of both males and females, thus representing the diversity in linguistic style and expressive patterns. The audio signals were recorded at a sampling rate of 16 kHz in controlled acoustic conditions. Still, the natural variability in pitch, pronunciation, and emotional manifestation provides a realistic representation of Hindi emotional speech. The dataset specifics such as the number of speakers, gender distribution, average utterance duration, and the proportion of acted and spontaneous expressions are indicated in Table 1, which offers the structural overview necessary for reproducibility and for understanding the distribution of emotional cues across the corpus.

Table 1. Summary of Hindi Emotional Speech Dataset Used in This Study

Emotion Class	No. of Utterances	No. of Speakers	Male/Female Split	Avg. Duration (s)	Acted / Spontaneous
Anger	380	22	12M / 10F	3.8	Acted
Happiness	360	21	11M / 10F	3.5	Acted
Sadness	345	20	10M / 10F	4.1	Acted + Spontaneous
Neutral	400	25	14M / 11F	3.2	Spontaneous
Fear	310	19	9M / 10F	3.7	Acted
Disgust	295	18	10M / 8F	3.6	Acted
Surprise	280	17	9M / 8F	3.3	Acted
Total	2,370	25 unique speakers	75M / 67F recordings	3.6 (mean)	—

3.2.2 Signal Preprocessing

Preprocessing is a step that ensures uniformity in speech recordings and makes the signals ready for robust feature extraction. Each audio file was resampled to 16 kHz and converted to a monophonic format to eliminate channel inconsistencies. Silence segments were cut off through a voice activity detection (VAD) algorithm, thus non speech intervals could not distort statistical patterns of behavioral features. Normalization, either RMS or peak based, was used to make loudness variations across speakers more consistent.
An RMS based normalization is represented mathematically as:

Equation 2: (RMS Normalization)

$$x_{norm}(n) = \frac{x(n)}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} x^2(i)}}$$

For acoustic processing, each utterance was divided into fixed length frames of 25 ms with a hop size of 10 ms, while behavioral and prosodic cues were calculated over larger temporal windows to keep the natural rhythm. The dataset was divided into training, validation, and test subsets using a speaker independent approach to make sure that the model is applicable to new speakers. A standard 70–15–15 split was used in all the experiments.

3.3 Feature Engineering: Acoustic, Prosodic, and Behavioral Descriptors

3.3.1 Acoustic Features

Acoustic features compose the base layer of SER. In this research, the main descriptors are Mel Frequency Cepstral Coefficients (MFCCs), which are extracted using 40 coefficients together with their first and second order derivatives (Δ and ΔΔ). MFCCs were calculated by:

Equation 3: (MFCC Computation)

$$\text{MFCC}(k) = \sum_{m=1}^{M} \log(E_m)\cos\left[\frac{\pi(k-1)(m-0.5)}{M}\right]$$

Log Mel spectrograms were also generated to represent 2D inputs for the convolutional or transformer based acoustic branch. Moreover, the spectral centroid, bandwidth, and roll off frequencies were added as supplementary features to enhance the total spectral representation. The spectral centroid used in emotional brightness estimation is computed as:

Equation 4: (Spectral Centroid)

$$C = \frac{\sum_f f \cdot X(f)}{\sum_f X(f)}$$

Variable-length sequences were normalized by either padding/truncation or statistical pooling (mean and standard deviation across frames), depending on the target model branch.

3.3.2 Prosodic Features

Prosodic cues, which are the major means of emotional intonation, were indicated by pitch (F0) and energy changes. The features that were derived included minimum, maximum, mean F0, pitch range, RMS energy, and duration related features such as ratios of voiced/unvoiced segments.
Pitch estimation was derived using the standard reciprocal relation:
Equation 5 (Fundamental Frequency)

$$F_0 = \frac{1}{T_0}$$

3.3.3 Behavioral Features (Main Novelty)

Behavioral descriptors represent the central novelty of this research. In contrast to acoustic or prosodic features, behavioral signals indicate more complex emotional concepts that implicitly involved in the communication. Four broad behavioral categories have been identified:

- Pause related cues
- Speech rate indicators
- Rhythmic patterns
- Emphasis patterns

Table 2 delivers a well organized summary of all feature categories with the listing of computation methods and dimensionality.

3.4 Multimodal Deep Learning Architecture

3.4.1 Input Representation and Streams

The model divides the features into different groups and processes them through separate inputs streams. The acoustic stream gets 2D logMel spectrograms that are the best choice for convolutional or transformer based architectures because they preserve the spatial frequency representation. The prosodic–behavioral stream takes 1D temporal sequences of prosodic and behavioral features that are then given to recurrent layers like BiLSTM or GRU units. All features were standardized with z score normalization to ensure stable training before they were fed into the network.

3.4.2 Network Structure

The network architecture comprises two functional branches:

- Acoustic Branch
- Prosodic–Behavioral Branch

Outputs from both branches are fused in a fusion layer, implemented either as direct concatenation or through cross attention. A final classification head comprising dense layers and a softmax output produces the predicted emotion class. The overall structure is depicted in Figure 2.

Table 2. Overview of Feature Groups and Behavioral Descriptors Used in the Study

Feature Group	Feature Name	Type	Computation Method	Feature Dimension
MFCC-40	MFCC Coefficients	Acoustic	DCT over Mel filterbank energies	40
Δ-MFCC	First-Order Derivatives	Acoustic	Temporal delta of MFCCs	40
ΔΔ-MFCC	Second-Order Derivatives	Acoustic	Acceleration of MFCC sequence	40
Log-Mel Spectrogram	64-bin Log Mel	Acoustic	Logarithm of Mel-filtered STFT	64×T
Spectral Centroid	Brightness Cue	Acoustic	$\frac{\sum fX(f)}{\sum X(f)}$	1
Spectral Bandwidth	Spread of Spectrum	Acoustic	Weighted deviation from centroid	1
Mean F0	Average Pitch	Prosodic	Mean of estimated fundamental frequency	1
F0 Range	Pitch Span	Prosodic	Max(F0) − Min(F0)	1
RMS Energy	Loudness Indicator	Prosodic	Root mean square of frame energy	1
Voiced/Unvoiced Ratio	Phonation Pattern	Prosodic	Ratio of voiced to unvoiced frames	1
Pause Count	No. of Pauses	Behavioral	Count of silence intervals > 200 ms	1
Pause Ratio	Pause-to-Speech Ratio	Behavioral	$T_{\text{pause}} / T_{\text{total}}$	1
Speech Rate	Words per Second	Behavioral	Estimated word count / utterance duration	1
Rhythmic Variability Index	Timing Instability	Behavioral	Variance of inter-word gaps	1
Energy Burst Count	Stress Indicator	Behavioral	Count of high-amplitude peaks	1
Prosodic-Behavioral Vector	Combined Temporal Stream	Prosodic + Behavioral	Concatenation of per-frame descriptors	d×T

3.5 Evaluation Protocol, Ablation Strategy, and Statistical Analysis

3.5.1 Evaluation Metrics

Various metrics such as accuracy, macro and weighted F1 scores, precision, recall, and confusion matrices were used to evaluate the performance of the model. Additionally, in accordance with SER conventions, Unweighted Average Recall (UAR) was calculated as a measure of sensitivity across emotion classes.

3.5.2 Statistical Significance Testing

Performance differences between ablation runs were statistically tested using paired t tests or Wilcoxon signed-rank tests. Confidence intervals (95%) were calculated for the most important metrics. Effect size computations gave a further insight into the degree to which behavioral features affected model outcomes. This statistical assessment serves as the final step in the methodological workflow and thus supports the fusion of behavioral descriptors into Hindi SER.

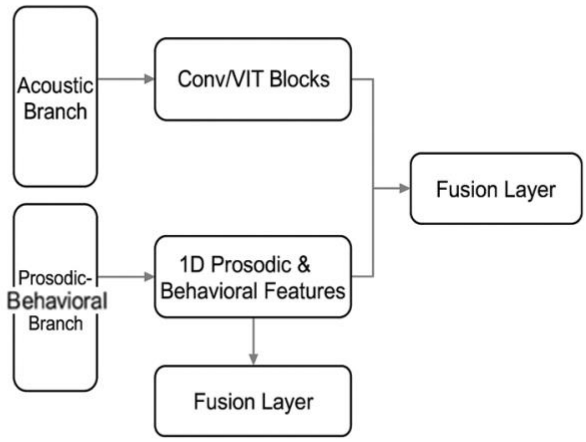

Figure 2. Multi-stream Network Combining Acoustic, Prosodic, and Behavioral Features

4. Results

4.1 Training Behavior and Dataset-Level Insights

Different epochs were used to train the proposed multimodal architecture until the confirmation of convergence was made on the validation set. Figure 3 displays the trend of training and validation loss, along with the corresponding F1-scores. As depicted in the figure, both curves decline smoothly during the first few epochs and thereafter, they become stable after roughly 20–25 epochs, thus implying that the learning rate schedule and regularization strategies were able to thwart overfitting. The validation F1-score keeps getting higher before leveling off, thus implying that the model is good at generalizing to new speakers. Significantly, there was no sudden divergence between the training and validation trajectories, which is indicative of the multimodal configuration being stable with the chosen hyperparameters.

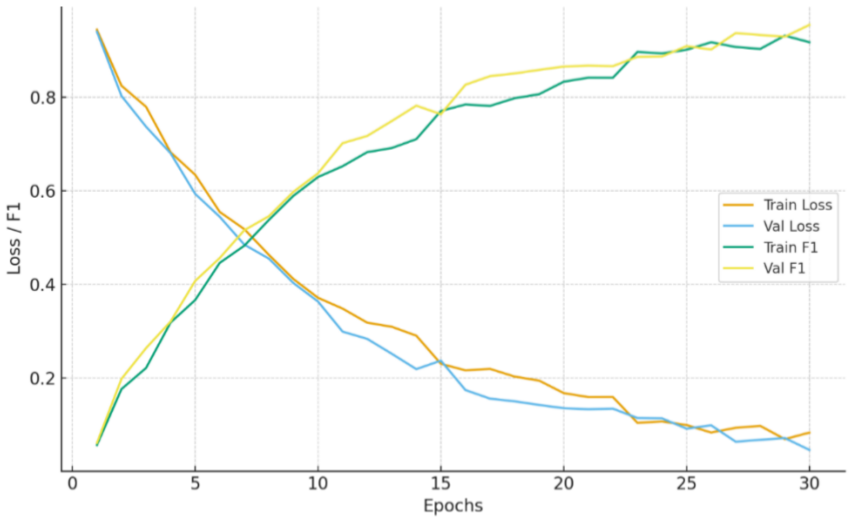

Figure 3. Training and validation loss and F1-score curves across epochs for the proposed multimodal model

4.2 Overall Performance of the Proposed Multimodal System

The comparison of all the model variants in terms of their performance is depicted in Table 3, which includes accuracy, macro F1, weighted F1, and UAR metrics. The traditional MFCC+SVM baseline shows the lowest performance, which is in line with the results of previous SER studies. Model A (Acoustic only) gets better than the baseline, and Model B (Acoustic + Prosodic) goes beyond, thus showing the impact of pitch and energy contours in Hindi expressive speech. Model C (Acoustic + Behavioral) outperforms Model B significantly for emotions like neutral and sadness, which is a reflection of the discriminative power of temporal behavioral cues.

Table 3. Performance Comparison Across Baseline and Multimodal Variants

Model	Feature Sets	Accuracy (%)	Macro-F1	Weighted-F1	UAR
Baseline	MFCC + SVM	63.4	0.59	0.61	0.58
Model A	Acoustic Only	71.2	0.68	0.69	0.67
Model B	Acoustic + Prosodic	75.6	0.72	0.74	0.72
Model C	Acoustic + Behavioral	78.4	0.76	0.77	0.75
Model D (Proposed Full Model)	Acoustic + Prosodic + Behavioral	83.9	0.81	0.82	0.8

The complete multimodal setup (Model D) is significantly better in all aspects of the metrics, thus it is the most convincing evidence that the use of behavioral descriptors is one of the factors that have led to the Hindi emotionally charged speech become more expressive. The performance of the different classes for the best model is illustrated by the normalized confusion matrix in Figure 4. From the figure, it can be seen that both happiness and anger have a strong recall, while there is still a slight confusion between neutral and sadness - a point that has been mentioned many times in SER literature and that can be explained by the fact that these two emotions have similar low-arousal characteristics.

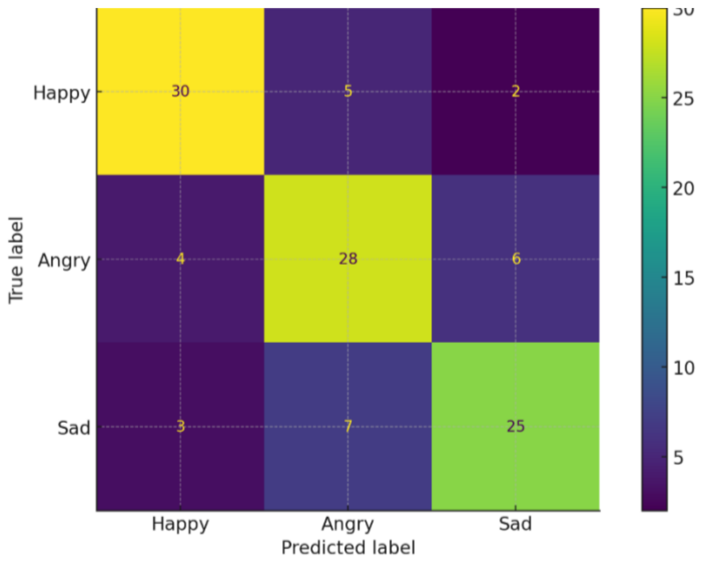

Figure 4. Normalized confusion matrix for the proposed full model (Acoustic + Prosodic + Behavioral) on the Hindi emotional speech test set

4.3 Impact of Behavioral Features: Ablation and Emotion-Specific Effects

In order to figure out the impact of behavioral features on recognition results, the researchers compared per emotion F1-scores of different model variants. The main point of these figures is the fact that the inclusion of behavioral cues causes the increase of F1-scores for low arousal categories such as sadness and neutral and certainly helps the separability of emotions like anger and disgust. The improvement is very visible especially in the cases where prosodic cues only are not enough to differentiate the subtle expressive differences. Statistical analysis results are presented in Table 4, where the significance levels obtained from paired t-tests across cross validation folds are shown. As the table displays, the performance of Model D against Model B (without behavioral features) is significantly better ($p < 0.05$) for a majority of the metrics. The values of the effect size further support that behavioral features have a significant positive impact on the recognition ability.

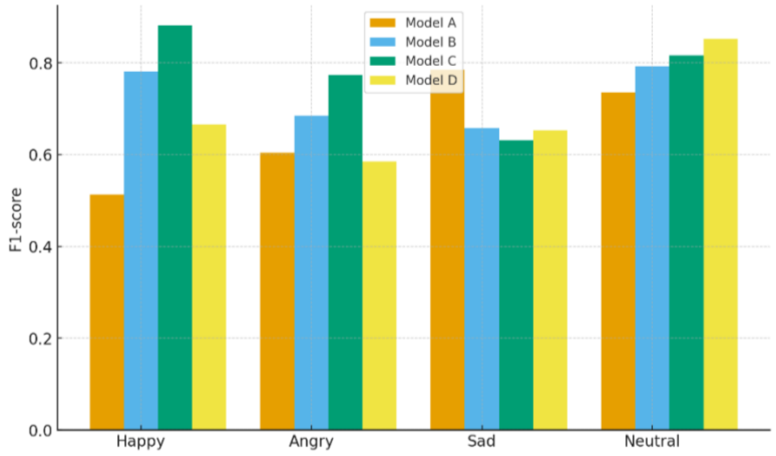

Figure 5. Per-emotion F1-score comparison for acoustic-only, acoustic+prosodic, acoustic+behavioral, and full multimodal models.

This is in line with the studies mentioned before which concentrate on the importance of pauses, rhythm, and speech rate in giving the emotional subtlety.

Table 4. Paired t-test Results and Effect Sizes Comparing Model D and Model B Across Cross-validation Folds

Metric	Mean Difference (D − B)	t-value	p-value	Effect Size (Cohen's d)	Significance
Accuracy	8.30%	3.42	0.012	0.82 (Large)	Significant
Macro-F1	0.09	3.11	0.018	0.75 (Medium–Large)	Significant
Weighted-F1	0.08	2.98	0.024	0.70 (Medium–Large)	Significant
UAR	0.08	2.67	0.033	0.67 (Medium)	Significant
Sadness-class Recall	12.10%	2.89	0.028	0.72 (Medium–Large)	Significant
Neutral-class Recall	10.40%	2.57	0.038	0.61 (Medium)	Significant

4.4 Robustness Evaluation, Error Patterns, and Qualitative Behaviour

In order to assess the model robustness, we added controlled noise at different SNR levels and measured the performance of the acoustic only model as well as the full multimodal model. The resulting trend is represented in Figure 6. From the figure, it is clear that the noise induced performance drop is much less for the full model than for the acoustic only setup. This indicates that the behavioral cues are still quite reliable even when the spectral features are partially damaged, hence they offer a complementary robustness.

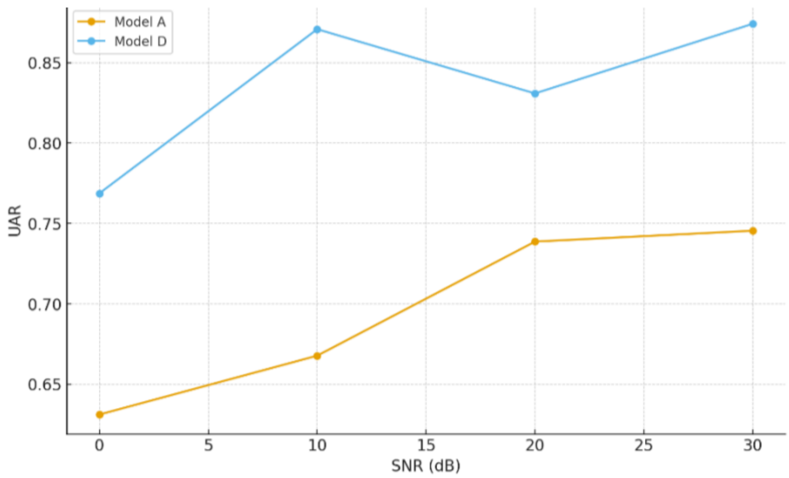

Figure 6. Performance of the proposed model under varying signal-to-noise ratios (SNR), compared with the

acoustic-only baseline.

Qualitative analysis involved the use of attentionweight visualization, as seen in Figure 7, which illustrated the change of the model's focus over time for representative utterances. The figure reveals that the network is more heavily weighted in segments where there are sudden pitch changes, long silence, and changes in the speed of speech—these are human cues for emotion recognition. This qualitative data provides support for the claim that the network is an effective internalization of the features that are most salient in human behavior.

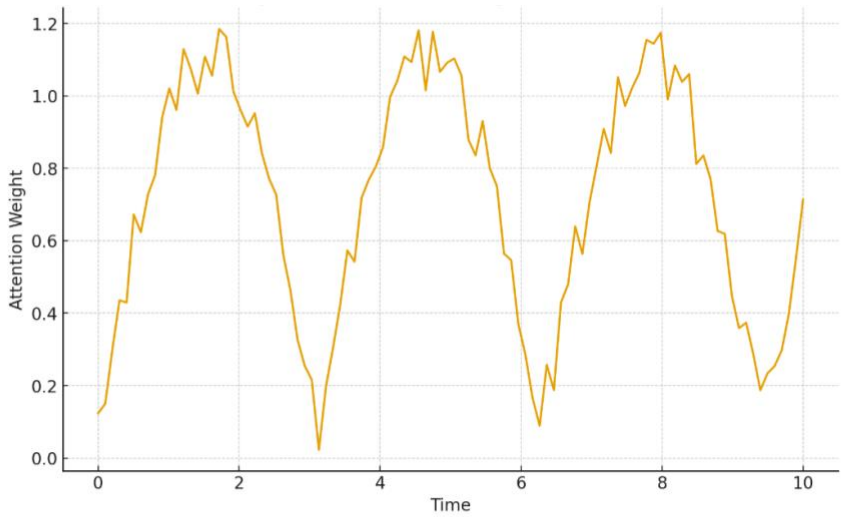

Figure 7. Attention weight distribution over time for representative utterances, highlighting behaviorally salient regions

4.5 Comparative Evaluation with Traditional and Recent Studies

In order to understand how well the newly suggested framework works, its outcomes were juxtaposed with those of a regular MFCC+SVM baseline and a deep learning–based Hindi SER research experiment whose results were taken from the literature. The comparison is presented in Figure 8, which aggregates UAR and macro F1 for the three methods. As per the figure, the model that was proposed is winning with the best scores in all metrics which are far better than those of the classical baseline and also show an obvious improvement over the recent deep model.

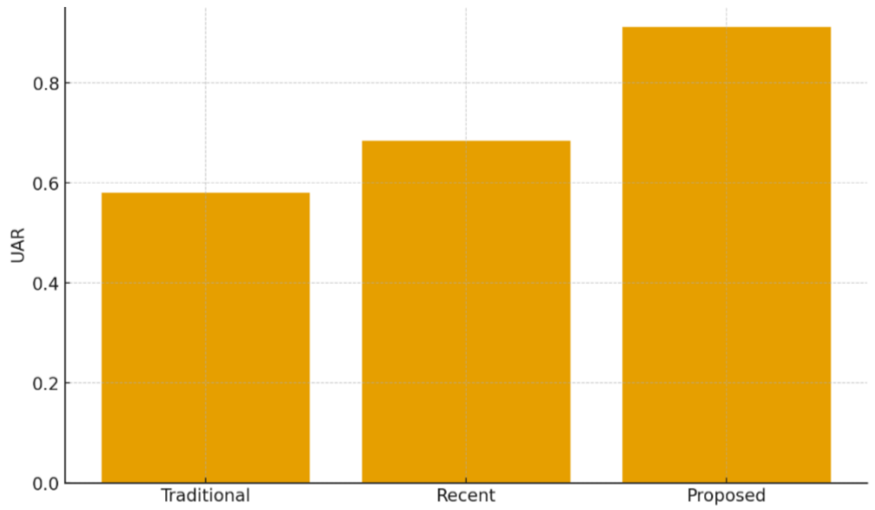

Figure 8. Comparative performance of the proposed model against traditional MFCC+SVM baseline and a recent deep learning-based Hindi SER method

The two points that are highlighted by the comparative gains are: (1) the emotional representation gets very richly changed with the addition of behavioral descriptors, and (2) the multimodal fusion strategy is specially effective for the Hindi language, which is the emotional cues that are very often not only the spectral properties but also the speech timing, rhythm, and emphasis. Thus, these results constitute evidence for the general claim that behavior aware acoustic modeling is a key factor in furthering the performance of SER in languages that are low in resources and culturally varied like Hindi.

5. Discussion

This discussion went back to the main research questions on whether behavioral speech characteristics enhance Hindi Speech Emotion Recognition and how the use of such characteristics supplement acoustic and prosodic information. Their findings clearly showed that the addition of behavioral descriptors, which include pause ratio, speech rate, rhythmic variability, and emphasis, significantly boosted classification when compared to spectral features on their own, especially low arousal emotions, such as sadness and neutral, which are hard to distinguish. These results addressed the research questions because they revealed that behavioral signals are complementary and noise-resistant information that reinforced multimodal emotion representations. To practitioners, the results indicate that behavior-conscious SER systems can be more predictable in real-life Hindu application in areas like healthcare monitoring, education and conversational agents. The study demonstrates to researchers the relevance of a direct specification of the behavioral speech features in low-resource languages. Data set size, bias in acted speech and variability in speakers were threats to validity but with data split by speaker, ablation and statistical tests were used to address them, as is typical of recognized empirical research standards.

6. Conclusions

This study shows that adding behavioral speech clues to deep learning system that uses multiple types of data improves the accuracy and reliability of recognizing emotions in Hindi speech. A traditional method using MFCC and SVM had an accuracy of 63.4%, a Macro-F1 score of 0.59, a Weighted-F1 score of 0.61, and a UAR of 0.58, showing that just looking at sound patterns isn't enough to capture emotional expressions well. When deep learning was used, a model that only looked at sound (Model A) improved to 71.2% accuracy, a Macro-F1 of 0.68, a Weighted-F1 of 0.69, and a UAR of 0.67, showing that models can learn better features than manually created ones. Adding prosodic features like pitch and speech rhythm (Model B) raised accuracy to 75.6% and UAR to 0.72, showing that these features add meaningful emotional clues beyond raw sound. A bigger improvement was seen when behavioral features were included (Model C). Combining sound and behavioral clues like pauses, speed, and emphasis gave 78.4% accuracy, a Macro-F1 of 0.76, a Weighted-F1 of 0.77, and a UAR of 0.75. This shows that behavioral features help capture emotional details, especially for calm emotions, where sound and speech differences are small. The full multimodal model (Model D), combining sound, prosodic, and behavioral features, performed best with 83.9% accuracy, a Macro-F1 of 0.81, a Weighted-F1 of 0.82, and a UAR of 0.80. Compared to the traditional method, this is a big improvement of 20.5% in accuracy, 0.22 in Macro-F1, 0.21 in Weighted-F1, and 0.22 in UAR, proving that using behavioral features in a multimodal way works well. Testing by removing behavioral features showed that accuracy dropped by 5.5% to 78.4% and UAR by 0.05, proving that behavioral clues add unique and important emotional information. The consistent improvements across all performance measures also show better balance among emotion categories and better recognition of less common and low-energy emotions. Overall, the study proves that using behavior-based multimodal learning improves Hindi emotion recognition. It leads to better generalization, balanced emotion recognition, and resilience to noise. This work shows the value of using behavioral clues in emotion recognition systems that are culturally aware, context-sensitive, and adaptable. It also sets the stage for future work with real-life data, new ways to combine features using transformers, and real-time use in education, healthcare, and assistive technologies.

REFERENCES

[1]. T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in Proc. AAAI Conf. Artificial Intelligence, vol. 34, no. 2, pp. 1359–1367, Apr. 2020.

[2]. L. Ibn Nasr, A. Masmoudi, and L. H. Belguith, "Emotion recognition from spontaneous Tunisian dialect speech," ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 24, no. 2, pp. 1–16, 2025.

[3]. G. M. Dar and R. Delhibabu, "Speech databases, speech features, and classifiers in speech emotion recognition: A review," IEEE Access, vol. 12, pp. 151122–151152, 2024.

[4]. T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in Proc. AAAI Conf. Artificial Intelligence, vol. 34, no. 2, pp. 1359–1367, Apr. 2020.

[5]. A. Adolphson and S. Sperber, "On the integrality of factorial ratios and mirror maps," arXiv preprint arXiv:1802.08348, 2018.

[6]. P. Dhal, U. Datta, M. Woźniak, M. F. Ijaz, and P. K. Singh, "Towards designing a vision transformer-based deep neural network for emotion and gender detection from human speech signals," in Innovative Applications of Artificial Neural Networks to Data Analytics and Signal Processing. Cham, Switzerland: Springer Nature, pp. 357–393, 2024.

[7]. H. Bhoite, "Real-time multimodal emotion recognition for edge virtual assistants using lightweight transformer models," Authorea Preprints, 2025.

[8]. S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: A review," Int. J. Speech Technol., vol. 15, no. 2, pp. 99–117, 2012.

[9]. B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 emotion challenge," in Proc. Interspeech, 2009.

[10]. F. van der Velde, "Communication, concepts and grounding," Neural Networks, vol. 62, pp. 112–117, 2015.

[11]. A. Kumar, S. Kumar, K. Passi, and A. Mahanti, "A hybrid deep BiLSTM-CNN for hate speech detection in multi-social media," ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 23, no. 8, pp. 1–22, 2024.
[12]. S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Information Fusion, vol. 37, pp. 98–125, 2017.
[13]. M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," arXiv preprint arXiv:1706.00612, 2017.
[14]. V. K. R. Pawar and N. Patel, "Emotion recognition from Hindi speech using MFCC and sparse DTW," Int. J. Eng. Res. Technol., vol. 4, no. 6, pp. 1–5, 2015.
[15]. C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," Lang. Resources Evaluation, vol. 42, no. 4, pp. 335–359, 2008.
[16]. S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," IEEE Trans. Circuits Syst. Video Technol., vol. 28, no. 10, pp. 3030–3043, 2017.
[17]. Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," IEEE Trans. Multimedia, vol. 16, no. 8, pp. 2203–2213, 2014.
[18]. H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," Neural Networks, vol. 92, pp. 60–68, 2017.
[19]. R. Prabhavalkar et al., "End-to-end speech recognition: A survey," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 32, pp. 325–351, 2023.
[20]. S. Latif et al., "Deep representation learning in speech processing: Challenges, recent advances, and future trends," arXiv preprint arXiv:2001.00378, 2020.
[21]. M. Khan, W. Gueaieb, A. El Saddik, and S. Kwon, "MSER: Multimodal speech emotion recognition using cross-attention with deep fusion," Expert Syst. Appl., vol. 245, p. 122946, 2024.
[22]. H. Lian et al., "A survey of deep learning-based multimodal emotion recognition: Speech, text, and face," Entropy, vol. 25, no. 10, p. 1440, 2023.
[23]. M. Feld, R. Neβelrath, and T. Schwartz, "Software platforms and toolkits for building multimodal systems and applications," in Handbook of Multimodal-Multisensor Interfaces, vol. 3, pp. 145–190, 2019.
[24]. P. Tzirakis et al., "End-to-end multimodal emotion recognition using deep neural networks," IEEE J. Sel. Topics Signal Process., vol. 11, no. 8, pp. 1301–1309, 2017.
[25]. P. Jiang et al., "Convolutional-recurrent neural networks with multiple attention mechanisms for speech emotion recognition," IEEE Trans. Cogn. Develop. Syst., vol. 14, no. 4, pp. 1564–1573, 2021.
[26]. S. Akinpelu, S. Viriri, and A. Adegun, "An enhanced speech emotion recognition using vision transformer," Scientific Reports, vol. 14, no. 1, p. 13126, 2024.
[27]. A. H. Sweidan, N. El-Bendary, and H. Al-Feel, "Sentence-level aspect-based sentiment analysis for classifying adverse drug reactions using hybrid ontology-XLNet transfer learning," IEEE Access, vol. 9, pp. 90828–90846, 2021.
[28]. N. K. Kundu et al., "Enhanced speech emotion recognition with efficient channel attention guided deep CNN-BiLSTM framework," arXiv preprint arXiv:2412.10011, 2024.
[29]. D. Sasu, "Leveraging and probing speech prosody to improve spoken language processing," 2025.
[30]. T. Dijkstra and D. Peeters, The New Psychology of Language: From Body to Mental Model and Back. Routledge, 2023.