Barun Singh Bisht, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Beyond Words: Emotion-Aware Automatic Speech Recognition Using Prosodic Feature Fusion

Barun Singh Bisht[1], Gurpreet Kaur[2], Gurpreet Kohli[3]
HMR Institute of Technology and Management
Affiliated to Guru Gobind Singh Indraprastha University, New Delhi, India

Abstract- Automatic Speech Recognition (ASR) systems achieved significant progress in converting speech into text; however, most existing approaches focused primarily on lexical accuracy and overlooked the emotional context present in human speech. This limitation reduced the effectiveness of ASR in applications that required natural and expressive interaction. In this study, a pipeline was developed for preserving emotional information from speech through prosodic feature analysis. Raw audio was processed through extraction, enhancement, alignment, and feature analysis stages to generate a structured dataset with synchronized annotations. Two models, a Support Vector Machine (SVM) and a Bidirectional Long Short-Term Memory (BiLSTM) network, were trained on six emotion classes, namely happy, sad, neutral, angry, fear, and curious, to evaluate their performance in emotion recognition. The experimental results showed that while the SVM provided a reasonable baseline, reaching 50 percent accuracy at 1000 samples per emotion, the BiLSTM model achieved higher accuracy of 69 percent under the same conditions, owing to its ability to capture temporal dependencies in speech. These findings highlight the importance of prosodic features and sequential modelling for developing more expressive and context-aware speech recognition systems.

Keywords: Speech emotion recognition; Prosodic feature fusion; Bidirectional Long Short-Term Memory.

I. INTRODUCTION

Automatic Speech Recognition has advanced considerably in recent years with the adoption of deep learning techniques and large scale datasets [1]. Modern systems can accurately convert spoken language into text and are commonly employed in virtual assistants, automated transcription systems, and voice driven user interfaces. However, most existing ASR systems focus primarily on recognizing spoken words and often ignore an important aspect of human communication, namely emotion.

In natural conversation, speech conveys both linguistic and emotional information. While the textual content represents what is being said, prosodic features such as pitch, intensity, and speaking rate reflect how it is said. These features play a critical role in conveying emotions such as happiness, anger, and curiosity, since the same sentence can express different meanings depending on the speaker's tone and delivery. Traditional ASR systems fail to capture these variations, which results in outputs that lack emotional context and reduces the overall quality of the interaction.

This study addressed this limitation by proposing a pipeline that preserves emotional information from speech through prosodic feature analysis. The collected audio data were processed through extraction, enhancement, alignment, and prosodic analysis stages to generate a structured dataset.

Two models were then trained on this dataset, a Support Vector Machine and a Bidirectional Long Short-Term Memory network, and were evaluated using six emotion classes, namely happy, sad, neutral, angry, fear, and curious, to analyze their ability to capture emotional patterns in speech. The results demonstrated that while traditional models such as SVM provide a baseline for classification, deep learning approaches such as BiLSTM perform considerably better, particularly as dataset size increases, which highlights the importance of temporal modeling in speech based emotion recognition and contributes toward more expressive and context aware ASR systems.

Historical Context of Emotion-Aware Speech Processing

Speech processing technologies evolved considerably over time, driven by the need to improve human machine interaction. Early ASR

© July 2026 Barun Singh Bisht, This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

systems, developed during the mid twentieth century, were primarily rule based and limited to recognizing isolated words, and often failed to cope with variations in speech such as different accents, speaking rates, and vocal tone, which limited their effectiveness in real world deployment. The introduction of statistical methods, particularly Hidden Markov Models, marked a major advancement during the 1980s and 1990s, since these models enabled the processing of continuous speech by capturing temporal variation in audio signals; however, their primary focus remained on identifying linguistic content, with little emphasis on the expressive aspects of speech.

Simultaneously, research on speech based emotion recognition progressed independently. Early approaches relied on handcrafted acoustic features and traditional classifiers, whereas more recent methods employed deep learning models, including bidirectional recurrent architectures [2], to capture temporal and contextual patterns. Despite these developments, the integration of emotion recognition into ASR systems remains a challenge, and this gap motivates the approach presented in this study, which jointly considers the linguistic and emotional aspects of speech.

Definitions and Key Terms

Understanding the fundamental concepts used in this study is important for interpreting the proposed approach and its results. Automatic Speech Recognition denotes the task of transforming spoken language into written text through computational and machine learning models; traditional ASR systems primarily focus on identifying phonemes and words from audio signals, often overlooking emotional and contextual aspects of speech. Prosodic features are non lexical elements of speech that contribute to its expressive nature, including pitch, intensity, speaking rate, and rhythm, and these features play a crucial role in conveying emotion, emphasis, and speaker intent.

Emotion recognition in speech entails determining a speaker's emotional state by analyzing acoustic and prosodic characteristics of the speech signal and classifying it into predefined emotional categories such as happy, sad, angry, or neutral. The Support Vector Machine is a supervised machine learning algorithm commonly used for classification and regression that identifies an optimal decision boundary, or hyperplane, separating classes in the feature space, and is effective for smaller datasets but may face limitations with complex, sequential data such as speech signals. The Bidirectional Long Short-Term Memory network is a recurrent architecture that processes a sequence in both forward and backward directions [2], allowing it to capture temporal dependencies that static models cannot. Feature fusion refers to combining multiple feature types, such as acoustic and prosodic features, into a single representation, which improves model performance by simultaneously capturing different aspects of speech.

Existing Work in Emotion-Aware Speech Systems

A substantial body of research has addressed speech based emotion recognition, with modeling approaches progressing from traditional machine learning techniques to more sophisticated deep learning architectures [5]. Initial work concentrated on extracting low level acoustic features, such as Mel Frequency Cepstral Coefficients, pitch, and energy, and feeding these into classifiers such as Support Vector Machines, k nearest neighbors, and decision trees. While these methods laid the groundwork for automatic emotion classification, their accuracy remained constrained because they struggled to capture the sequential and temporal structure inherent in speech signals. Cross corpus studies further showed that the performance of such classifiers varies considerably across datasets and recording conditions, which underlines the need for more robust and generalizable modeling strategies [3].

With the advancement of deep learning, researchers began adopting models such as Convolutional Neural Networks and Long Short-Term Memory networks, which improved the modeling of sequential dependencies in speech. The categorical emotion labels used across this body of work, including happiness, sadness, anger, and fear, draw on the broader theory of basic universal emotions established in psychological research [4]. Despite

this progress, there remains a need for efficient approaches that integrate emotion recognition within the speech processing pipeline while primarily relying on speech signals rather than auxiliary modalities.

Role of Prosody in Speech Understanding

Prosody plays a fundamental role in human communication by adding expressive and contextual information to spoken language. While words convey the literal meaning of a sentence, prosodic features such as pitch, intensity, speaking rate, and rhythm determine how a message is interpreted. Incorporating prosodic features into speech processing systems enables a more comprehensive understanding of spoken language, since analyzing patterns in pitch, energy, and temporal dynamics makes it possible to detect emotional variation and enhance the quality of interaction between humans and machines. This is particularly important in applications such as conversational agents, virtual assistants, and assistive technologies, where understanding both what is said and how it is said is essential for effective communication.

Research Gap and Motivation

Despite significant advances in both Automatic Speech Recognition and speech based emotion recognition, a clear gap still exists in integrating these two domains effectively. Most modern ASR systems are designed to maximize transcription accuracy and focus primarily on recognizing linguistic content, and as a result tend to ignore the emotional and expressive aspects of speech that are essential for understanding human communication in a natural way. This gap motivated the present study, which aims to preserve emotional information within an ASR oriented processing pipeline.

Research Objectives

The primary objective of this study was to develop an approach for preserving emotional information in speech by leveraging prosodic features within a structured processing pipeline, thereby enhancing traditional speech processing through the incorporation of emotional context. To achieve this, the study aimed to design and implement a pipeline for collecting, processing, and aligning speech data with corresponding annotations; to extract and analyze prosodic features such as pitch, intensity, and temporal variation from speech signals; to develop emotion classification models using both traditional machine learning and deep learning approaches; to implement and evaluate Support Vector Machine and Bidirectional Long Short-Term Memory models for emotion recognition; to analyze the impact of dataset size on model performance and learning capability; to compare the effectiveness of non sequential and sequential models in capturing emotional patterns in speech; and to contribute toward the development of more expressive and context aware speech processing systems.

II. METHODOLOGY

Data Collection

The data collection phase focused on gathering emotionally rich speech data from real world sources to capture natural variation in human expression. Instead of relying solely on controlled recordings, this study used audio extracted from story narrations, web series, and movies, since these sources contain naturally occurring emotional speech and make the dataset more representative of real world scenarios than scripted or laboratory recorded data. Story narrations provided clear and expressive speech patterns, web series captured conversational and spontaneous emotion, and movies contributed a wide range of intense and subtle emotional expression. Audio clips were extracted from dialogues in which emotions were clearly identifiable, which ensured relevance for the emotion recognition task.

Dataset Source and Emotion Categories

The extracted audio samples were categorized into six emotion classes: happy, representing joyful and energetic expression; sad, representing low energy and subdued tone; neutral, representing emotionally balanced speech; angry, representing high intensity and aggressive tone; fear, representing anxious or tense expression; and curious, representing a questioning or inquisitive tone. These categories were selected to cover a broad emotional spectrum, consistent with established taxonomies of basic

emotion [4], while maintaining clarity in classification.

Manual Annotation and Data Organization

All collected audio samples were manually annotated to ensure high quality labeling. Each audio clip was listened to carefully, assigned an appropriate emotion label based on tone and context, transcribed into text, and marked with timestamps for precise alignment between speech and labels. Manual annotation was preferred over automated labeling because it provided better accuracy in capturing the subtle emotional nuances present in natural speech. The resulting dataset was organized into audio files in the .wav format, corresponding text transcriptions, and annotation files in CSV or TextGrid format containing emotion labels with timestamps, with each file following a consistent naming convention to maintain alignment across the audio, text, and annotations.

This data collection strategy captured real world emotional variability and natural prosodic patterns, which improved model generalization and provided diverse speaking styles and contexts. Nonetheless, the process involved certain challenges, including background noise in movie and web series clips, overlapping dialogue in some scenes, subjectivity in manual emotion labeling, and imbalance in emotion distribution, all of which were addressed during the preprocessing and model training stages described below.

Audio Processing

After the dataset was collected and annotated, the raw audio was processed to improve its quality and ensure consistency across samples. Because the data were extracted from stories, web series, and movies, they often contained background noise, varying volume levels, and irregular pauses, which made preprocessing essential before feature extraction and model training.

Noise reduction was applied first, since clips obtained from movies and web series often included background music, environmental sound, and overlapping dialogue that could negatively affect feature extraction; spectral subtraction techniques, together with low pass and high pass filtering, were used to remove irrelevant frequency components and isolate the speaker's voice [8]. Audio normalization was then applied to standardize amplitude levels and ensure uniform loudness across clips of varying recording conditions, allowing the model to focus on emotional patterns rather than volume differences. Silence removal followed, using Voice Activity Detection to trim silent segments below a defined energy threshold and produce compact, meaningful audio data. Long clips were subsequently divided into smaller segments based on sentence level boundaries, phrase level pauses, and annotation timestamps, which ensured that each segment contained a clear and consistent emotional expression. Finally, because audio clips from different sources had different sampling rates, all files were resampled to a fixed rate of 16 kHz to maintain consistency and reduce computational complexity.

The complete audio processing pipeline therefore proceeded through extraction from the source, noise reduction, normalization, silence removal, segmentation, and resampling, in that order, which ensured that the final audio data were clean, consistent, and ready for feature extraction. The main challenges encountered at this stage were removing noise without losing emotional cues, handling overlapping speech in dialogue, preserving prosodic features during filtering, and managing variability in recording quality; these were addressed through careful tuning of the preprocessing techniques to balance noise removal against the preservation of emotional information.

Feature Extraction

Feature extraction is among the most critical stages in the proposed system, since it transforms raw audio signals into meaningful numerical representations that can be used for recognizing emotion. In this study, both prosodic and acoustic features were extracted to capture not only what was being said but also how it was being said.

The prosodic features extracted included pitch, or fundamental frequency, which represents the tone of speech and is often higher for emotions such as

happiness or excitement and lower for sadness; intensity, or energy, which measures loudness and tends to be higher for angry or excited speech and softer for sad speech; speaking rate, which reflects the speed of delivery, with faster speech often indicating excitement or anger and slower speech indicating sadness or hesitation; and duration, which captures the length of phonemes, words, and pauses and helps identify emotional state. In addition to these prosodic cues, acoustic features representing the spectral properties of speech were extracted, including Mel Frequency Cepstral Coefficients, which capture the perceptual characteristics of the audio signal and are widely used in speech processing; the spectral centroid, which indicates the center of mass of the frequency spectrum; spectral flux, which measures changes in the spectrum over time; and the zero crossing rate, which represents how often the signal changes sign and helps distinguish voiced from unvoiced sounds.

Feature extraction was performed using the Librosa Python library for Mel Frequency Cepstral Coefficients, spectral features, and zero crossing rate [9], and using Praat for pitch, intensity, and duration related features [6]. To improve model performance, the prosodic and acoustic features were concatenated into a unified feature vector, which allowed the fused representation to capture both emotional and linguistic characteristics and enhanced the models' ability to distinguish between similar emotions. Before the features were supplied to the models, min max scaling was applied to bring values into a fixed range between zero and one, and z score normalization was used to center the data around the mean with unit variance, which prevented certain features from dominating others because of scale differences.

The two models required different input formats: the SVM used fixed length, flattened feature vectors, whereas the BiLSTM used sequential, time series feature inputs that preserved temporal dependencies in speech. The principal challenges in this stage were the effect of noise on pitch and intensity estimation, variability in speech across speakers, difficulty in capturing subtle emotional differences, and the high dimensionality of the combined feature set, all of which were mitigated through careful preprocessing and feature selection.

Model Training

Two models were used for emotion classification in this study, a Support Vector Machine and a Bidirectional Long Short-Term Memory network, with the objective of comparing a traditional machine learning approach against a deep learning model capable of capturing temporal dependencies in speech. Before training, the dataset was divided into training and testing sets using an 80:20 split, the extracted features were scaled and normalized to ensure uniform input, and the emotion category labels were encoded into numerical form; features for the BiLSTM were arranged as sequential, time series data, while features for the SVM were converted into fixed length vectors.

The SVM was used as a baseline model to evaluate the effectiveness of traditional classification techniques, implemented using the scikit-learn library [10]. It works by identifying an optimal hyperplane that separates the different emotion classes within a high dimensional feature space, configured with a Radial Basis Function kernel, a regularization parameter that controls margin flexibility, and a gamma parameter that defines the influence of individual data points. During training, feature vectors combining prosodic and acoustic information were supplied as input, and the model learned boundaries between emotion classes, with multi class classification handled using One-vs-One or One-vs-Rest strategies. The SVM was effective for smaller datasets, was less computationally expensive, and performed well when features were clearly separable, but it could not capture sequential or temporal information and its performance decreased as the dataset grew larger and more complex.

The BiLSTM model, implemented using TensorFlow and Keras [11], [12], was used to capture temporal dependencies in speech signals, making it more suitable for emotion recognition. It processes the input sequence in both forward and backward directions, which allows it to understand context from both past and future frames [2]. The

architecture consisted of an input layer accepting sequential feature vectors, one or more BiLSTM layers capturing temporal dependencies, a dropout layer to prevent overfitting, a dense fully connected layer, and a softmax layer producing a probability distribution over the emotion classes. The model was trained using categorical cross entropy loss and the Adam optimizer, with batch size tuned experimentally and the number of epochs selected based on convergence. The BiLSTM captured long term dependencies and utilized context from both directions, performed well on larger datasets, and was better at recognizing subtle emotional variation than the SVM, though at the cost of greater computational expense, a larger training data requirement, and longer training time.

To improve performance, the key hyperparameters of each model were tuned: for the SVM, the regularization parameter, gamma, and kernel type; and for the BiLSTM, the number of layers, hidden units, learning rate, and dropout rate, using grid search or manual tuning to determine the optimal values. The complete training workflow proceeded through input feature extraction, feature scaling and formatting, the train test split, training of both the SVM and BiLSTM models, hyperparameter tuning, and model validation. The main challenges encountered during training were overfitting in the deep learning model, data imbalance across emotion classes, the high computational requirement of the BiLSTM, and the difficulty of tuning hyperparameters, which were addressed through regularization and dataset balancing techniques.

Evaluation Framework

The evaluation phase assessed the performance of the proposed models in accurately recognizing emotion from speech. Both the SVM and BiLSTM models were evaluated using standard classification metrics, namely accuracy, precision, recall, and F1-score, to ensure a fair and comprehensive comparison; these metrics are defined in detail in Section 3.2. A confusion matrix was also used to visualize model performance in detail, showing correct and incorrect predictions for each emotion class, helping to identify which emotions were often misclassified, and providing insight beyond overall accuracy, particularly for confusion between similar emotions such as sad versus neutral and fear versus curious.

The dataset was split into a training set comprising 80 percent of the samples and a testing set comprising the remaining 20 percent, which ensured that each model was tested on unseen data and provided a realistic measure of generalization capability; cross validation could optionally be applied to further improve reliability. The SVM and BiLSTM models were then compared on the basis of overall accuracy, performance across individual emotion classes, ability to handle sequential data, and computational efficiency, in order to determine which model was more suitable for emotion aware speech recognition. Class imbalance, the similarity between certain emotions, speaker variability, and limited dataset size were identified as the main challenges affecting this evaluation, and were addressed through careful dataset design and the use of balanced evaluation metrics.

III. ANALYSIS

Experimental Setup

This section describes the dataset, implementation environment, and configuration used to train and evaluate the models. The audio samples were collected from stories, web series, and movies and grouped into six emotion classes, namely happy, sad, neutral, angry, fear, and curious, with a total of 1000 samples per emotion. As in the methodology, the dataset was split into 80 percent for training and 20 percent for testing to evaluate model generalization. The system was implemented in Python, using Librosa for feature extraction of Mel Frequency Cepstral Coefficients, pitch, and intensity [9]; NumPy and Pandas for data handling; scikit-learn for the SVM model [10]; and TensorFlow with Keras for the BiLSTM model [11], [12]. Demucs was used for source separation and Audacity for manual audio cleaning during preprocessing [7], [8].

Model training was carried out on a workstation equipped with an Intel Core i5 processor, 32 GB of RAM, and an NVIDIA GeForce RTX 4070 Super GPU

with 12 GB of GDDR6X memory. The experimental procedure proceeded through prosodic and acoustic feature extraction, data preparation including normalization, the train test split, and sequential formatting for the BiLSTM; training of both the SVM baseline and the BiLSTM deep learning model; and evaluation using accuracy, precision, recall, F1-score, and the confusion matrix.

Performance Metrics

To evaluate the effectiveness of the proposed emotion aware speech recognition system, four standard classification metrics were used, which together provide a comprehensive understanding of how well the models recognize emotion.

Accuracy is the proportion of correctly classified samples relative to the total number of samples and provides a general measure of overall model performance, computed as follows.

Accuracy = TP + TNTP + TN + FP + FN

Precision measures the proportion of correctly predicted instances of an emotion among all instances predicted as that emotion, and indicates the model's reliability in predicting a particular emotion without producing false alarms.

Precision = TPTP + FP

Recall measures the proportion of correctly predicted instances of an emotion relative to all actual instances of that emotion, and shows how well the model identifies all occurrences of a given emotion.

Recall = TPTP + FN

The F1-score is the harmonic mean of precision and recall and provides a balanced measure of performance, which is particularly useful when classes are imbalanced.

F1-Score = 2 × Precision × RecallPrecision + Recall

A confusion matrix was additionally used to show the correct and incorrect predictions for each emotion class in tabular form, which helped identify which emotions were most often misclassified and provided insight into each model's weaknesses; the confusion matrices for the SVM and BiLSTM models are discussed in Section 4.4.

IV. RESULTS

Results of the Support Vector Machine Model

The Support Vector Machine model served as a baseline for evaluating traditional machine learning methods for emotion recognition. It was trained on dataset sizes ranging from 100 to 1000 samples per emotion to examine the effect of data size on performance, as summarized in Table 1.

Table 1. SVM accuracy across dataset sizes

Model	Samples per Emotion	Accuracy (%)
SVM	100	23
SVM	250	34
SVM	500	41
SVM	750	46
SVM	1000	50

As shown in Table 1, accuracy improved steadily as the dataset size increased for each emotion, remaining moderate for smaller datasets of 100 to 250 samples per emotion and reaching 50 percent at 1000 samples per emotion. Emotions with distinct prosodic features, such as anger and happiness, were classified more accurately than subtle emotions, such as sadness and neutrality, and this pattern is consistent with the SVM's inability to capture temporal dependencies, which limits its performance on sequential speech features.

Table 2. SVM precision, recall, and F1-score at 1000 samples per emotion

Emotion Class	Precision	Recall	F1-Score
Happy	0.70	0.68	0.69
Sad	0.52	0.50	0.51
Neutral	0.50	0.48	0.49
Angry	0.75	0.72	0.73
Fear	0.45	0.43	0.44
Curious	0.42	0.40	0.41

The per class metrics in Table 2 confirm that the SVM provided reasonable baseline performance and improved clearly as dataset size increased, and that it was more effective for emotions with distinct prosodic features than for subtle or similar emotions. This finding motivates the use of a model capable of temporal or sequential learning, namely the BiLSTM, discussed next.

Results of the BiLSTM Model

The BiLSTM model leverages temporal dependencies in speech to improve emotion recognition and was trained on datasets ranging from 250 to 1000 samples per emotion, as summarized in Table 3.

Table 3. BiLSTM accuracy across dataset sizes

Model	Samples per Emotion	Accuracy (%)
BiLSTM	250	43
BiLSTM	500	57
BiLSTM	750	63
BiLSTM	1000	69

The BiLSTM consistently outperformed the SVM at every comparable dataset size, with accuracy improving considerably as more data became available, which demonstrates the advantage of sequential modeling. The model captured subtle emotional nuances more effectively than the SVM, particularly for the sad, neutral, fear, and curious classes.

Table 4. BiLSTM precision, recall, and F1-score at 1000 samples per emotion

Emotion Class	Precision	Recall	F1-Score
Happy	0.88	0.85	0.86
Sad	0.70	0.68	0.69
Neutral	0.68	0.65	0.66
Angry	0.90	0.87	0.88
Fear	0.65	0.63	0.64
Curious	0.62	0.60	0.61

As shown in Table 4, the BiLSTM was more effective than the SVM across all dataset sizes, since sequential learning allowed it to recognize temporal patterns in speech and accuracy gains were more pronounced with larger datasets, which demonstrates the model's scalability. These results confirm the importance of prosodic feature fusion for emotion aware ASR.

Comparative Analysis

This section presents a direct comparison between the SVM and BiLSTM models in terms of accuracy, scalability with dataset size, and ability to capture temporal dependencies in speech, summarized in Table 5.

Table 5. Accuracy comparison across dataset sizes

Model	Samples per Emotion	Accuracy (%)
SVM	100	23
SVM	250	34
SVM	500	41
SVM	750	46
SVM	1000	50
BiLSTM	250	43
BiLSTM	500	57
BiLSTM	750	63
BiLSTM	1000	69

The BiLSTM consistently outperformed the SVM for all comparable dataset sizes. The SVM improved gradually but plateaued near 50 percent accuracy, whereas the BiLSTM continued improving up to 69 percent at 1000 samples per emotion, and the difference between the two models was more pronounced for smaller datasets, which highlights the ability of the BiLSTM to leverage sequential information even with limited data. Table 6 summarizes the relative strengths and weaknesses of the two models across several criteria.

Table 6. Model strengths and weaknesses

Criteria	SVM	BiLSTM
Accuracy	Moderate	High
Temporal dependencies	Not captured	Captured (both directions)
Dataset size sensitivity	Limited improvement with larger data	Improves significantly with data
Performance on subtle emotions	Poor (e.g., sad vs. neutral)	Better (captures subtle patterns)
Computational cost	Low	High
Scalability	Limited	High

These results indicate that sequential learning is central to the BiLSTM's advantage, since its ability to capture temporal patterns allows it to outperform the SVM across all emotions, although both models benefit from the pitch, intensity, and speaking rate features described in Section 2.3. Accuracy for both models improved with larger datasets per emotion, but the SVM saturated earlier than the BiLSTM, and these practical implications suggest that the BiLSTM is better suited to emotion aware ASR systems, particularly in real world applications requiring nuanced emotion recognition.

Overall, the BiLSTM clearly outperformed the SVM in terms of accuracy, temporal modeling, and handling of subtle emotional differences, while the SVM can still serve as a lightweight baseline for smaller datasets or resource limited scenarios; this comparative analysis justifies the choice of the BiLSTM as the primary model for emotion aware ASR in this study.

Confusion Matrix Analysis

Confusion matrices provide detailed insight into how each model classifies emotion. To understand the impact of dataset size, the SVM and BiLSTM performance were compared across 100, 250, 500, 750, and 1000 samples per emotion, summarized in Table 7 and Table 8.

Table 7. SVM confusion matrix observations across dataset sizes

Samples per Emotion	Accuracy (%)	Key Observations
100	23	High misclassification for sad versus neutral and fear versus curious; distinct emotions such as angry and happy were recognized more reliably.
250	34	Misclassifications were slightly reduced, though subtle emotions remained unclear.
500	41	Errors decreased further as the model began learning prosodic patterns more effectively.
750	46	Misclassification rates dropped, especially for fear versus curious, though sad versus neutral confusion persisted.
1000	50	Lowest error rate achieved; subtle emotions still posed challenges, while angry and happy reached near perfect classification.

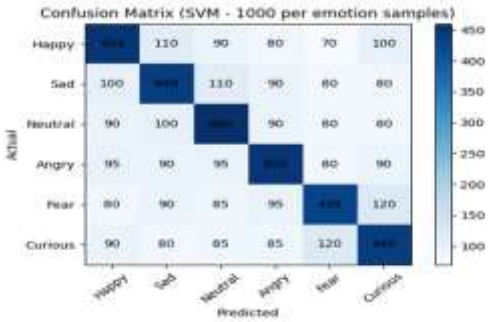

Figure 1. SVM confusion matrix at 1000 samples per emotion

As dataset size per emotion increased, the SVM improved gradually, but its reliance on static features meant it continued to struggle with temporal and prosodic patterns, which limited its maximum achievable accuracy; misclassifications decreased

steadily but remained prominent for subtle emotions.

Table 8. BiLSTM confusion matrix observations across dataset sizes

Samples per Emotion	Accuracy (%)	Key Observations
250	43	Already better than the SVM at the same dataset size; misclassifications mainly involved sad versus neutral and fear versus curious.
500	57	Errors decreased significantly as the BiLSTM began capturing sequential prosodic patterns.
750	63	Most misclassifications were reduced, and subtle emotions were better distinguished.
1000	69	Minimal errors remained; angry, happy, and sad were almost all correctly classified, with only slight confusion between fear and curious.

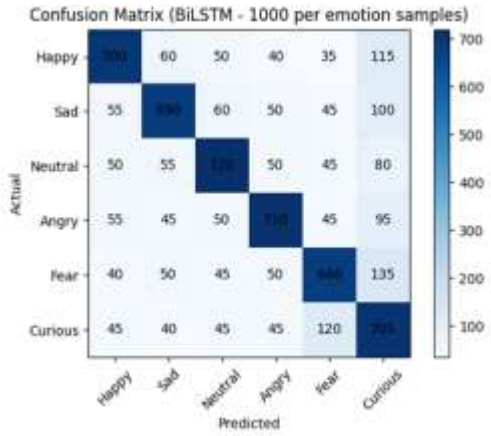

Figure 2. BiLSTM confusion matrix at 1000 samples per emotion

The BiLSTM reduced misclassification more rapidly than the SVM as dataset size increased, since sequential modeling captured temporal patterns that helped differentiate subtle emotions; the largest improvement was observed between 250 and 500 samples per emotion, indicating that both data availability and model capacity matter for performance.

Table 9. Comparative misclassification patterns across dataset sizes

Samples per Emotion	SVM Accuracy (%)	BiLSTM Accuracy (%)	Key Misclassifications
100	23	--	Sad vs. neutral; fear vs. curious
250	34	43	Same pattern, though BiLSTM produced fewer errors
500	41	57	BiLSTM reduced subtle emotion confusion
750	46	63	Most BiLSTM misclassifications removed
1000	50	69	BiLSTM minimal errors; SVM still struggled with subtle emotions

Overall, accuracy improved steadily with more samples per emotion for both models, though the BiLSTM reduced misclassification faster because of its temporal modeling capability and was significantly better at distinguishing sad from neutral and fear from curious than the SVM. At 1000 samples per emotion, the BiLSTM achieved 69 percent accuracy with minimal confusion, whereas the SVM reached only 50 percent with persistent errors on subtle emotions.

V. DISCUSSION

Interpretation of Findings

The results presented in Section 4 show that incorporating prosodic features allowed both models to capture emotional information beyond plain lexical content, and that the choice of model architecture had a substantial effect on how well this information was exploited. The BiLSTM consistently outperformed the SVM across all dataset sizes, confirming that temporal modeling is important for recognizing emotional patterns that unfold over the course of an utterance rather than being captured in a single, static feature vector. Dataset size also

played a clear role in performance: both models improved as the number of samples per emotion increased, but the BiLSTM scaled more effectively with larger datasets because of its capacity to model sequential dependencies.

Pitch, intensity, and speaking rate proved critical for distinguishing emotional states, and subtle emotions such as sad versus neutral and fear versus curious were recognized more accurately when these features were combined with sequential modeling rather than used with a purely static classifier. Taken together, these findings indicate that a complete pipeline spanning data collection, annotation, audio processing, feature extraction, and model training and evaluation can achieve emotion aware ASR with proper feature representation and sequence modeling.

Limitations

Although this study demonstrated the effectiveness of emotion aware ASR, several limitations should be noted. The SVM struggled with temporal and subtle emotional patterns, which limited its ability to classify nuanced emotions accurately even with larger datasets, while the BiLSTM, although more effective, required higher computational resources and longer training times than traditional models. The dataset itself, drawn from stories, web series, and movies, may not fully represent natural, spontaneous speech encountered in everyday conversation, and only six emotion classes were considered, which do not cover the full spectrum of human emotion.

Models trained on scripted or acted speech may not generalize well to real world environments in which background noise, accents, and speaking styles vary considerably, and emotion recognition remains inherently subjective, which introduces the possibility of annotation bias despite careful manual labeling. Finally, while prosodic features such as pitch, intensity, and speaking rate were used, other modalities such as voice timbre, additional spectral features, and facial expression were not included in the training data, and the pipeline does not currently handle multilingual speech, which limits its applicability in diverse linguistic settings. In summary, although the proposed system produced promising results, these limitations highlight the continuing need for larger and more diverse datasets, multimodal feature integration, and more robust modeling techniques to improve the real world applicability of emotion aware ASR.

Future Work

Several directions could extend this research. Human emotion is far more nuanced than the six categories considered here and often exists in combinations or gradations, such as frustration mixed with anger or relief mixed with happiness; future work could therefore explore hierarchical emotion modeling, in which emotions are first classified into primary categories and then into finer subcategories, allowing ASR systems to become more emotionally aware and closer to human perception. Because speech alone conveys only part of human emotional expression, integrating audio with visual cues such as facial expression, lip movement, and body gesture through multimodal fusion techniques, for example combining BiLSTM audio models with convolutional networks for visual features or attention based transformers for cross modal interaction, could significantly enhance the recognition of subtle emotions that are difficult to distinguish from audio alone.

Because current models, particularly the BiLSTM, require substantial computational resources and may not be suitable for real time use, future work could focus on model optimization techniques such as pruning, quantization, or knowledge distillation to enable deployment on edge devices such as smartphones or embedded systems, which would allow real time emotion recognition to enhance virtual assistants, call centers, and interactive artificial intelligence more broadly. The current dataset, based on stories, web series, and movies, may also limit generalization to natural speech, so future research should incorporate larger and more diverse datasets, including real conversations, interviews, and public speeches, and apply data augmentation techniques such as pitch shifting, speed variation, noise addition, and voice conversion to expand the dataset artificially and help the models learn more robust prosodic patterns under varied acoustic conditions.

Pretrained speech models such as Wav2Vec2, Whisper, and HuBERT capture rich audio representations and can reduce the need for large labeled datasets; fine tuning such models with emotion specific layers could improve performance, particularly for subtle emotions, and future work could also explore cross lingual transfer learning to recognize emotion across different languages, dialects, and accents.

Because speech emotion is often influenced by context, such as prior conversation history, speaker personality, or situational circumstance, incorporating contextual information, for example through sequence to sequence context modeling, transformers with attention over previous utterances, or speaker embeddings, could improve recognition in ambiguous cases where tone varies with sarcasm, stress, or excitement. Finally, because emotion recognition is currently treated as a separate module performed after feature extraction, future systems could explore end to end joint modeling in which transcription and emotion recognition are learned simultaneously, which could reduce error propagation, improve efficiency, and enable emotion preserving transcription in real time applications.

VI. CONCLUSION

This study investigated the integration of emotional information into Automatic Speech Recognition through prosodic feature fusion, with the goal of preserving the emotional content of speech alongside accurate transcription. A complete processing pipeline was designed, spanning data collection from stories, web series, and movies, manual annotation, audio preprocessing, prosodic and acoustic feature extraction, and the training and evaluation of two classification models.

The Bidirectional Long Short-Term Memory model consistently outperformed the Support Vector Machine across all dataset sizes, reaching 69 percent accuracy at 1000 samples per emotion compared with 50 percent for the SVM, and both models improved as dataset size increased, though the BiLSTM scaled more effectively because of its ability to model temporal dependencies. Pitch, intensity, and speaking rate proved critical for distinguishing emotional states, with subtle emotions such as sad versus neutral and fear versus curious recognized more accurately when these prosodic features were combined with sequential modeling. The novelty of this work lies in its use of naturally occurring, real world emotional speech drawn from unscripted and scripted media rather than laboratory recordings, combined with a systematic comparison of static and sequential classifiers under varying data availability.

At the same time, the study is limited by its reliance on a six class emotion taxonomy, a monolingual dataset, and prosodic features alone without complementary modalities such as facial expression, and by the higher computational cost associated with the BiLSTM. Overall, the study confirms that integrating prosodic features with deep sequential models such as BiLSTM can meaningfully enhance the ability of ASR systems to recognize and preserve emotion in speech, motivating further work on larger, more diverse, and multimodal datasets.

VI. DECLARATIONS

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not for profit sectors.

Data Availability

The audio dataset and annotations generated and analyzed during this study are available from the corresponding author upon reasonable request.

Author Contributions

All authors contributed to the conception and design of the study, data collection and annotation, model development, and analysis of results, and all authors read and approved the final manuscript.

REFERENCES

1. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015.
2. A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," in Proc. International Joint Conference on Neural Networks (IJCNN), 2005, pp. 2047–2052.
3. B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, and G. Rigoll, "Cross-corpora acoustic emotion recognition: Variances and strategies," IEEE Transactions on Affective Computing, vol. 1, no. 2, pp. 119–131, Jul.–Dec. 2010.
4. P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," Journal of Personality and Social Psychology, vol. 17, no. 2, pp. 124–129, 1971.
5. M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, vol. 44, no. 3, pp. 572–587, Mar. 2011.
6. P. Boersma, "Praat: Doing phonetics by computer," Version 6.3.14, 2023. [Online]. Available: https://www.praat.org
7. Audacity Team, "Audacity: Free, open-source, cross-platform audio software," Version 3.3.3, 2024. [Online]. Available: https://www.audacityteam.org
8. Facebook AI Research, "Demucs: Deep extractor for music sources," 2021. [Online]. Available: https://github.com/facebookresearch/demucs
9. S. McFee et al., "librosa: Audio and music signal analysis in Python," in Proc. Python in Science Conference (SciPy), 2015.
10. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
11. M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: https://www.tensorflow.org
12. Keras Team, "Keras: The Python deep learning API," 2015. [Online]. Available: https://keras.io