Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

Integrating Automatic Speech Recognition and Emotion Detection: A Conformer-XGBoost Framework for Human-Centered Speech Systems

Mohan Bikram K C.[1], Smita Adhikari[2], Tara Bahadur Thapa[3]

[1]Department of Computer Engineering, Gandaki College of Engineering and Science, Pokhara, Nepal.

[2]Assitant Professor, Department of Electronics and Computer Engineering, Pashchimanchal Campus, Pokhara, Nepal.

[3]Assitant Professor, Department of Computer Engineering, Gandaki College of Engineering and Science, Pokhara, Nepal.

Email: [1]kcmohan64@gmail.com, [2]adsmita1@gmail.com, [3]tara@gces.edu.np

Abstract

Advanced speech technology pushes human-machine interaction to a new frontier. Most of the models address this either as a matter of speech-to-text transcription or emotion detection. By integrating an XGBoost-driven emotion classification component with a Conformer-based speech recognition system, an integrated solution has been developed. It will, therefore, strive to transcribe spoken utterances and estimate the emotional condition of the speaker with as much accuracy as possible to improve context-sensitive interaction. The transcription process combines large, multilingual speech corpora. A Conformer architecture captures both short- and long-range temporal dependencies. In this regard, an error rate of 0.322 words and 0.146 characters was achieved in transcription. For emotion recognition, several emotional speech datasets were collected, and various acoustic features were extracted under noisy conditions. Using an XGBoost model, 86.58% accuracy in emotion detection was attained. These results demonstrate the feasibility of integrating speech transcription with emotion recognition and form a basis for the further development of more human-like, empathic, and adaptive voice systems.

DOI: https://doi.org/10.36548/jaicn.2025.4.003
Received: 10.10.2025, received in revised form: 08.11.2025, accepted: 22.11.2025, published: 02.12.2025

Keywords: Automatic Speech Recognition, Speech Emotion Recognition, Conformer Architecture, XGBoost Classifier, Human Computer Interaction, Multimodal Speech Processing.

1. Introduction

During the last few years, both NLP and AI have experienced a sea change, bringing about an entirely new paradigm in the interaction between humans and technology. Today, integrated SER and ASR support language interfaces that are aligned with human dialogue and remain context-sensitive. Instead of isolated modules, unified pipelines based on shared signals are set replace them. The attention-based neural framework, specifically the Transformer model, has recently revolutionized sequence-to-sequence learning due to its parallelism without recurrence or convolutions, which is efficient and scales well (Vaswani et al. [1]). Among the domains, speech intelligibility and fast information exchange are of prime importance. To advance communication and information management, the integration of SER into speech-to-text systems is proposed in current research to ensure that the conducted analysis and reaction will be context sensitive. In other words, the aim is end-to-end processing-from capturing an audio signal to actionable output. Transcription and emotion analysis for meetings, education, and customer service converge into a single AI-powered solution:

ASR bridges spoken language to machine comprehension, allowing transcription services, voice assistants, and other forms of interactive tools, including a fixed interface to text-input components. High-accuracy ASR reduces miscommunication and enables support of real-time processing for audio applications beyond what has been possible until now. SER estimates the affective state in speech to frame intent and emotional context. Treated together, ASR and SER provide context-aware outputs that represent both semantic content and affect rather than mere text. This will take HCI toward the development of systems that are much more empathetic, far more understanding, and even more effective. Thus, responses can consider not just what is said but how it is expressed.

1.1 Problem Statement

The existing systems primarily treat Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER) as independent tasks due to limitations in speech technologies, which restricts their ability to understand the emotional context of speech and the information

conveyed. This division hinders the accurate transcription of empathy or contextual understanding, making it challenging to participate in real-life conversations with background noise, varying accents, and differing emotional intensities.

1.2 Objective

The main objective of the study is to create an integrated framework that includes a Speech Emotion Recognition model driven by XGBoost and an Automatic Speech Recognition system based on Conformer, to deliver precise transcriptions of spoken language while discerning the speaker's hidden emotions.

2. Related Work

Amodei et al. [2] presented Deep Speech 2, the end-to-end speech recognition system for English and Mandarin that employs deep neural networks instead of complicated hand-engineered pipelines. The rates of measurement errors, presently considered word and character error rates (WER and CER), were considerably improved compared to previous methods with both recurrent and convolutional architectures using the CTC loss. Through many optimization procedures executed by GPUs and data from hundreds of hours of speech for training, Deep Speech 2 was proven to be relatively effective in multilingual, noisy, and accented use-case scenarios, offering transcription at near-human levels of quality. This work thereby shows how powerful scalable deep learning architectures may be in fostering breakthroughs in automatic speech recognition.

Aouani and Ayed [3] designed a two stage SER system with feature extraction and classification. The extraction stage computes MFCCs, zero crossing rate, the Teager energy operator, and harmonic to noise ratio. An autoencoder compresses the feature vectors as an unsupervised dimensionality reduction step. An SVM then assigns emotion labels to the compact representation. Tests on the RML corpus show that autoencoder based fusion gives higher recognition accuracy than earlier methods. This setup separates representation learning from the final decision step.

Heafield [4] introduced KenLM, a library for fast n-gram language model queries with low time and memory cost. The library provides two back ends that target different constraints. PROBING uses a linear probing hash table to maximize speed. TRIE uses bit level packing and interpolation search to lower memory use. The two options expose a clear speed memory

trade off. Against SRILM and IRSTLM, KenLM yields up to 2.4x speedup and much lower memory use. The code is thread safe and open source and is integrated into Moses, cdec, and Joshua. KenLM is widely adopted in MT pipelines.

Kim and Kwak [5] proposed an SER framework that combines explainable AI with transfer learning. Speech waveforms are converted to spectrograms and pass through preprocessing that reduces uncertainty. The system applies a late fusion ensemble over pretrained audio models such as YAMNet and VGGish. The framework links model predictions with explanations while using features from large pretrained networks. The goal is to improve interpretability while retaining the benefits of transfer learning.

Kumar, Mahrishi, and Nawaz [6] examined speech sentiment analysis from a machine learning perspective. The review categorizes methods by feature types, including prosodic, spectral, and linguistic descriptors. It examines classifiers, including SVMs and NNs. The authors notably do comparative analyses across many datasets to identify strategies that exhibit consistent robustness under diverse settings. The survey summarizes common settings and highlights consistent patterns across corpora.

Sahu [7] addressed multimodal SER. The study computes eight acoustic features from the signal and uses them as inputs to several classifiers, including SVMs, decision trees, feedforward networks, and LSTMs. The pipeline also adds linguistic cues from text to reduce ambiguity that arises with acoustic-only evidence. The results indicate that multimodal input can reduce interpretive ambiguity relative to acoustic-only models. The comparative experiments in this study across a variety of models suggest that relatively simple, hand-crafted features, when combined with hybrid modalities, can sometimes outperform more complex deep learning baselines, complicating assumptions about the need for highly complex architectures for effective emotion recognition.

Sutskever, Vinyals, and Le [8] introduce a fully neural “sequence-to-sequence” architecture that utilizes one multilayer LSTM encoder to encode an input sequence into a fixed-dimensional vector, and a second multilayer LSTM decoder to decode the target sequence from that fixed-size embedding. They improve performance by reversing the input word order for the encoder, which makes it easier for the model to learn short-term dependencies.

Vijayakumar, Singh, and Mohanty [9] presented a real-time system that converts spoken language into text, summarizes it, and delivers the summary back as speech. The framework utilizes Deep Speech 2 for speech recognition and Abstract Meaning Representation graphs as a mechanism for producing abstractive summaries. Furthermore, it implements preprocessing steps, such as CTC loss and batch normalization, to improve recognition accuracy. The study emphasizes that the combination of speech recognition, text summarization, and text-to-speech will enable local communication support, from sender to receiver, more efficiently and with greater capability in a non-intrusive manner compared to other forms of communication.

3. Proposed Work

It proposes an integrated multimodal framework in which Speech Emotion Recognition and Automatic Speech Recognition techniques are combined in a way that offers significantly enhanced computer interaction based on both linguistic and emotional understanding. Besides generating very accurate transcriptions, the system is designed to detect the speaker's emotions.

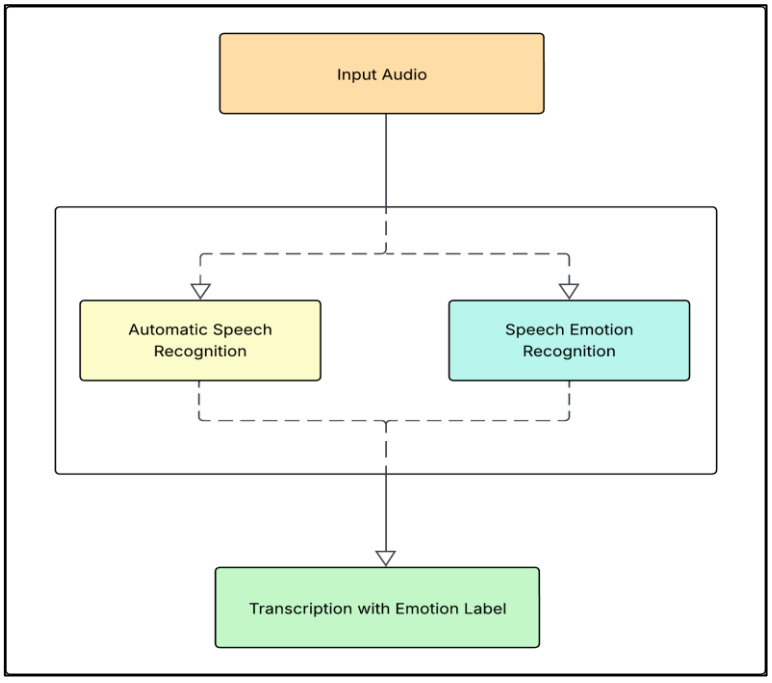

Figure 1. System Architecture

Figure 1 illustrates the integrated flow of the proposed system. Input audio is processed by Automatic Speech Recognition and Speech Emotion Recognition to generate a transcription that includes emotion labels.

3.1 Automatic Speech Recognition

3.1.1 Datasets

The Mozilla Common Voice dataset is a very large multilingual corpus with hundreds of hours of audio clips and their transcripts. More specifically, the corpus Common Voice Corpus 21.0 was employed. 90% of the 3,670-hour corpus, roughly 3,303 hours, was used for training, while the other 10%, approximately 367 hours, were used for validation. Recordings apart from this were collected using Mimic Recording Studio, which were used for fine-tuning to improve performance against accent variations.

3.1.2 Data Preprocessing

Then, Sox with multithreading converted the raw MP3 files into FLAC format featuring lossless audio and efficient processing. Acoustic features consisted of Mel Spectrogram transformations. Applications like frequency masking and time masking were used to enhance performance against environmental noise and different speaking rates, padding was applied to the input of the spectrogram, resulting in a final product of transcription that accommodates unequal sequence lengths.

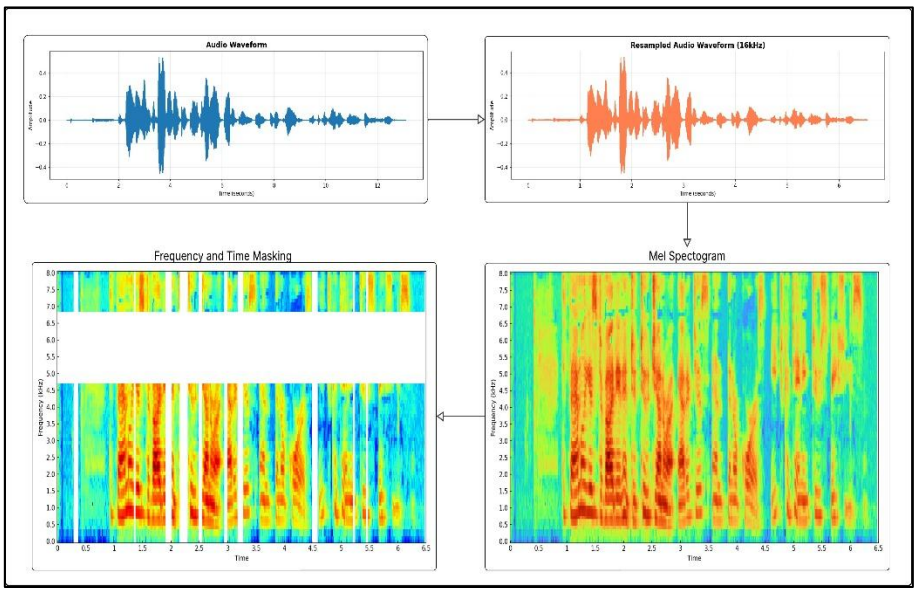

Figure 2. Audio Preprocessing Pipeline for Automatic Speech Recognition

Figure 2 describes the processing chain from the raw audio waveform to a resampled waveform and then, its corresponding spectrogram representation that will be used as input to the model.

3.1.3 Model Architecture

An original work by Gulati et al. [11] attempted to utilize an architecture that leveraged both local and global contextual information in speech. The encoder uses convolutional subsampling, followed by a stack of Conformer blocks combining feed-forward layers relative multi-head self-attention, and convolutional modules that capture short and long-term dependencies. This extends the Transformer to incorporate convolutional operations in each block. This helps to model local acoustic patterns, such as formant transitions and short-term spectral variations. Self-attention captures global context over the entire sequence, while the convolutional module focuses on fine-grained temporal details. The hybrid design allows it to represent both linguistic and acoustic cues far better than standard Transformer models for speech sequences. Finally, the decoder employs an LSTM that projects encoder outputs onto class logits for character-level predictions. Training is done with CTC, which stands for Connectionist Temporal Classification, a task that allows for alignment-free mapping between acoustic features and text sequences. It can model the complex temporal dynamics in speech and be consistent with arbitrary input lengths.

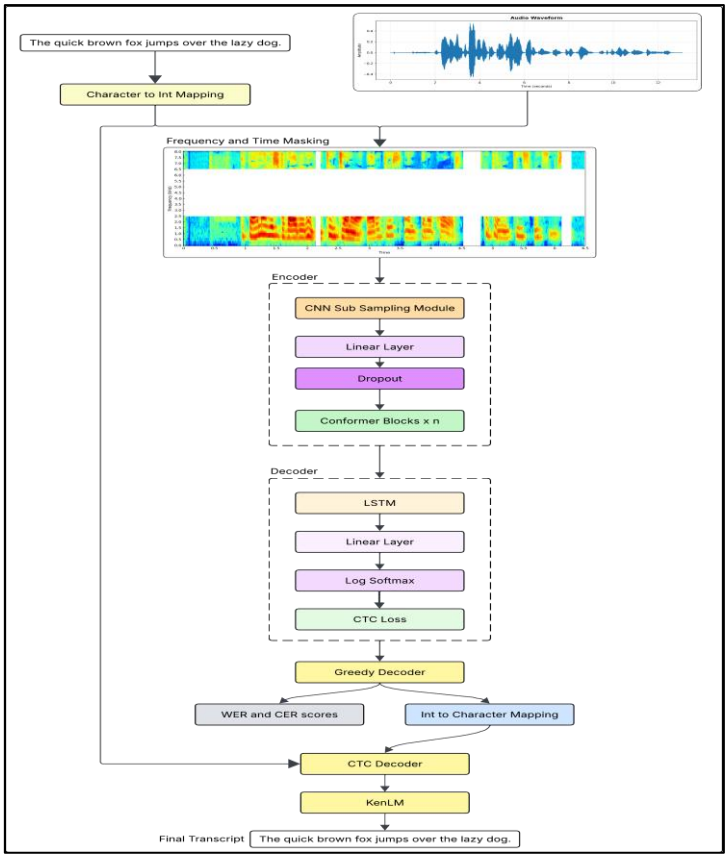

Figure 3. Architecture of Automatic Speech Recognition Model

Figure 3: An end-to-end ASR architecture; audio waveforms are processed to extract Mel-spectrogram features with augmentation using frequency masking and time masking, after which features go through a CNN subsampling encoder, followed by multiple Conformer blocks, and an LSTM decoder that uses CTC loss to predict the final transcript.

3.2 Speech Emotion Recognition

3.2.1 Datasets

Several datasets were combined in order to provide diversity within the expression of emotion. Among these are RAVDESS (Livingstone and Russo [12]), CREMA-D (Cao et al. [13]), TESS (Pichora-Fuller, and Dupuis [14]), SAVEE (Jackson and ul haq [15]), and ESD (Zhou et al. 16). These datasets consist of speech and audio-visual samples in .wav, .mp3, .mp4, and FLAC formats with transcripts covering anger, disgust, fear, happiness, sadness, surprise, and neutral. The data was then split into training and testing sets in the ratio of 80-20.

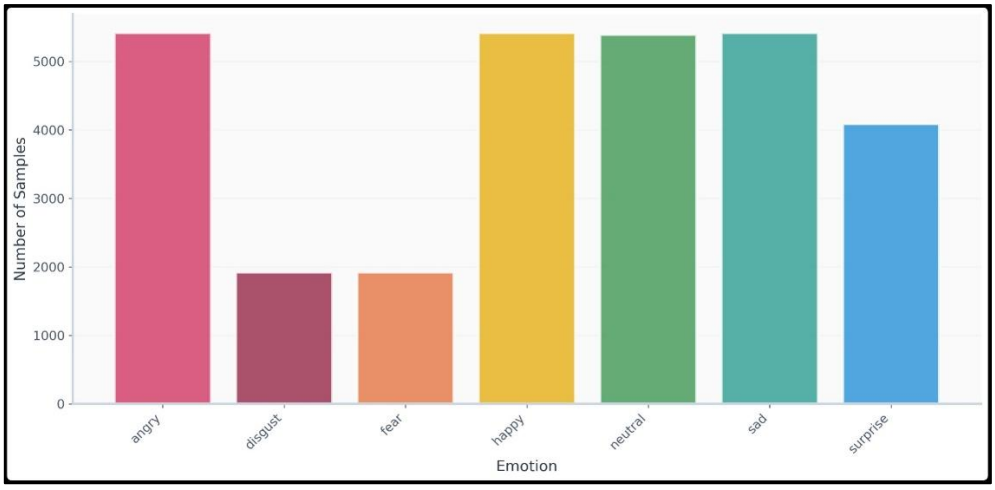

Figure 4. Distribution of Emotion Classes in the Dataset

Figure 4 presents the distribution of samples across seven emotion categories in the dataset. This figure indicates that angry, happy, neutral, and sad are the most frequent classes in both datasets, while disgust, fear, and surprise represent the classes which are underrepresented in these datasets.

3.2.2 Data Preprocessing

This speech audio was first converted to FLAC and resampled to 16 kHz so that all the files were brought to a similar standard of processing. The data augmentation techniques-which are complementary in nature, ensure that the models are robust and generalized. These

techniques included noise addition, time stretching, pitch shifting, and time shifting. Other complementary acoustic properties were accounted for through feature extraction, whereby MFCCs capture phonetic and timbral properties, the Chromagram captures the tonal and harmonic properties, while the Spectral Contrast captures fluctuations in energy across different frequency bands. To handle the problem of class imbalance that characterizes emotional categories, care was taken to ensure that the classes were kept proportional in their representation through stratified splitting of the data. During model training, sample weights were assigned to mitigate bias toward the dominant emotions. Lastly, after feature extraction, the features and their corresponding emotion labels were stored in an organized manner and used later in the training and testing of the models.

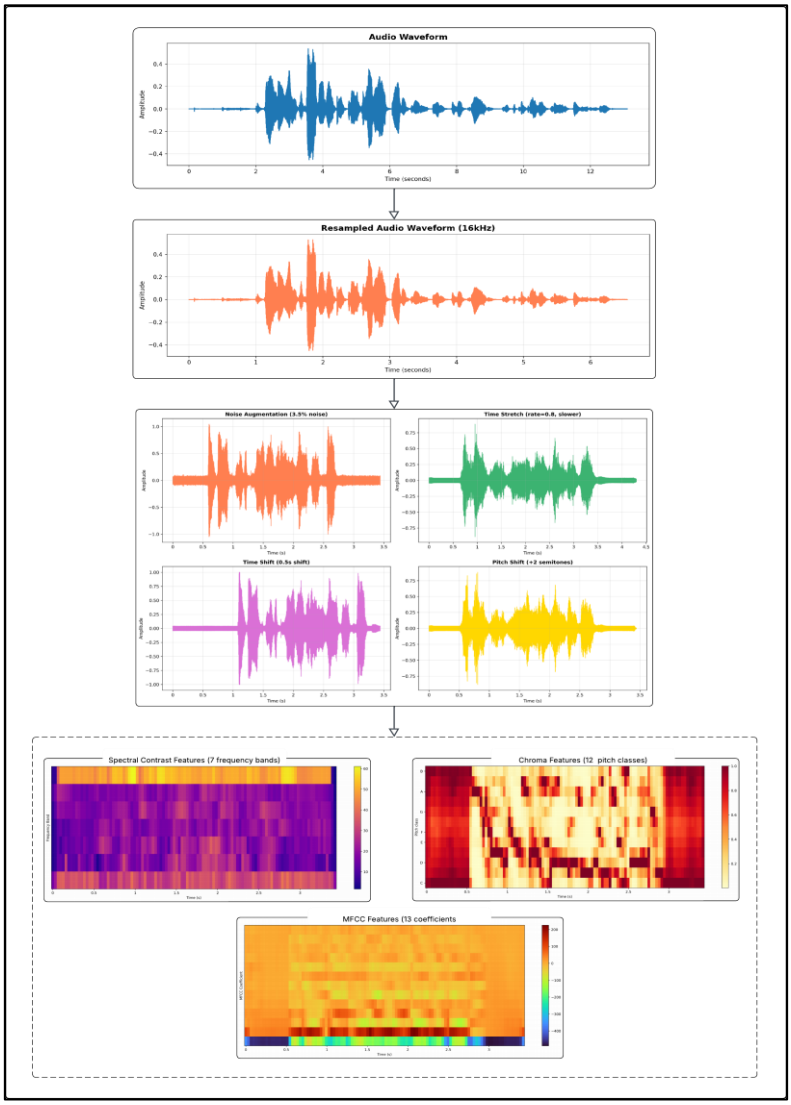

Figure 5. Audio Preprocessing and Feature Extraction Pipeline for Speech Emotion Recognition

Figure 5: Audio Preprocessing and Feature Extraction Pipeline for Speech Emotion Recognition Figure 5 depicts the full chain of audio pre-processing, going from raw audio waveform to resampling at 16kHz, passing through the data augmentation techniques of adding noise, time stretching, and pitch shifting, and finishing with the extraction of acoustic features using spectral contrast, chroma, and MFCC representation.

3.2.3 Model Architecture

A classic XGBoost classifier was trained on acoustic features such as MFCCs, chroma, and spectral contrast extracted from audio. These features have been previously standardized using z-score normalization, while their corresponding labels were encoded in order to prepare for multi-class classification. Z-score normalization is used to scale features consistently across all dimensions and hence stabilizes model training, keeping it away from biases related to features with larger numerical ranges. We have chosen the XGBoost algorithm for being computationally efficient and suitable for small to medium-sized datasets. We apply the following settings thereby enabling XGBoost to work with medium-sized datasets effectively during training without excessive resource demand. It is also robust against overfitting through mechanisms such as regularization and tree pruning that ensure stable generalization across classes. The application of XGBoost agrees with the nature of SER features, which are tabular representations of prosodic and spectral characteristics, benefiting from learning based on gradient boosting. Hyperparameter tuning was performed with grid search together with cross-validation to achieve the best performance of the model. Besides this, generated evaluation metrics included accuracy, precision, recall, F1-score, confusion matrices, and visualizations of class distributions, feature importance, and prediction confidence. For enhanced reliability downstream, the output probabilities of the classifier were post-hoc calibrated by isotonic regression to make the confidence values better reflect true likelihoods. This allows the integrated system to set thresholds on decisions more coherently and interpretably and enhances the reliability of emotion-driven responses in later modules. Quantification of prediction uncertainty was given as the entropy of the calibrated probability distribution, characterizing the dispersion of confidence across emotion classes. These uncertainty estimates defined operating points that balance precision and coverage whereby predictions with entropy greater than the threshold set were excluded from decisions downstream.

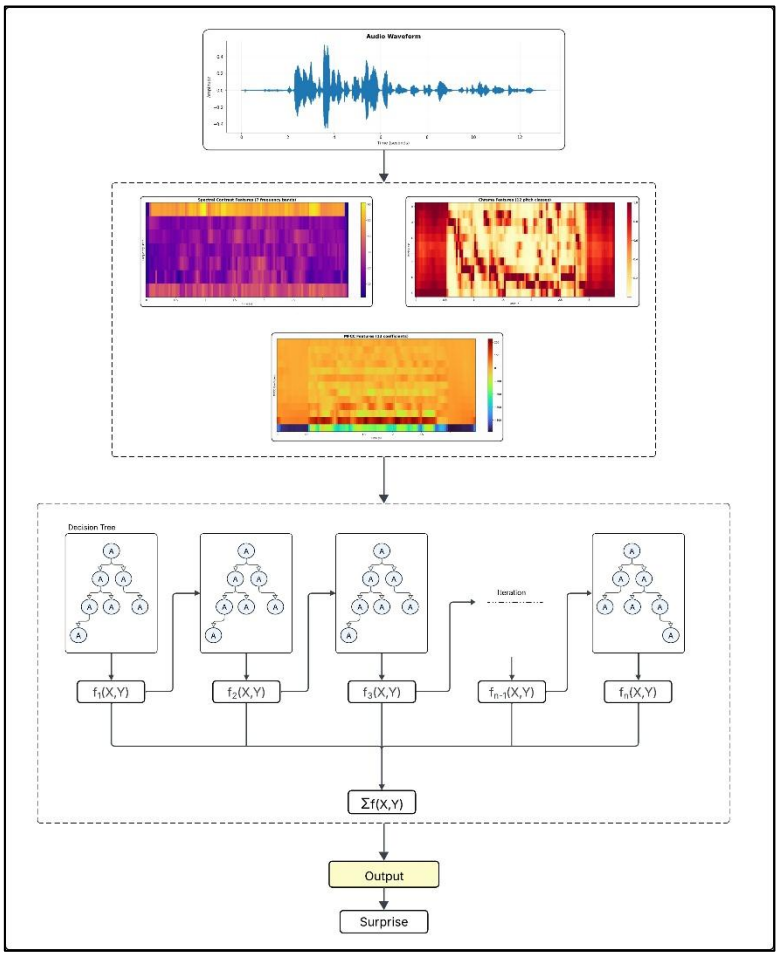

Figure 6. Architecture of Speech Emotion Recognition Model

Figure 6. Architecture of Speech Emotion Recognition Model Figure 6 illustrates an example XGBoost architecture for audio sentiment classification. In this example, augmented audio waveforms are converted into spectral contrast and chroma, MFCC features that are further processed by a number of decision trees whose outputs are summed up to produce the ultimate sentiment prediction.

4. Results and Discussion

4.1 Automatic Speech Recognition

Advanced speech technology has pushed the boundaries of human-machine interaction even further. Instead of viewing it as a separate set of speech transcription and emotion detection systems, most models consider the process as either speech-to-text transcription or emotion detection. Here, we integrate a Conformer-based speech recognition module along with an XGBoost-driven emotion detection system. The performance of the ASR model was

measured using Word Error Rate (WER) and Character Error Rate (CER), both of which assess transcription accuracy by comparing predicted outputs against ground truth references.

The convergence of the training process was smooth through epochs, with both the training and validation losses consistently dropping. The model reached its best performance at epoch 34, with no further improvement afterward. Thus, early stopping due to this reason eventually halted the training at epoch 38. At the optimal checkpoint—epoch 34—the model attained a validation character error rate (CER) of 0.146 and a word error rate (WER) of 0.322.

A WER of 0.322 indicates that roughly every third word was incorrectly transcribed. This represents competitive performance for medium-sized datasets using the Conformer architecture. The CER of 0.146 means most character predictions were correct, signaling proper alignment between predicted and reference sequences. Given this, the model achieved dependable quality in transcription. The early stopping rule stopped training before overfitting and preserved generalization on unseen data.

The observed accuracy results from the Conformer architecture that integrates convolution and self-attention. The convolutional modules model the short-term dependencies and local phonetic features, reducing substitution and insertion errors. The attention layers model long-range contextual relationships necessary for speech-to-text mapping. The architecture processes the local and global properties of speech signals through this dual mechanism.

The real-time factor of the model was about 0.02 during inference, with the capability of processing 60 seconds of audio in 2-3 seconds. While loading the model, the peak memory usage was 3.2 GB, whereas for making inferences, it used 200 to 300 MB. On the CPU, with 8 cores and 16 threads, the average usage stays between 2-4%, thus maintaining low latency. These numbers demonstrate how real-time speech recognition applications can be quickly and easily deployed to production.

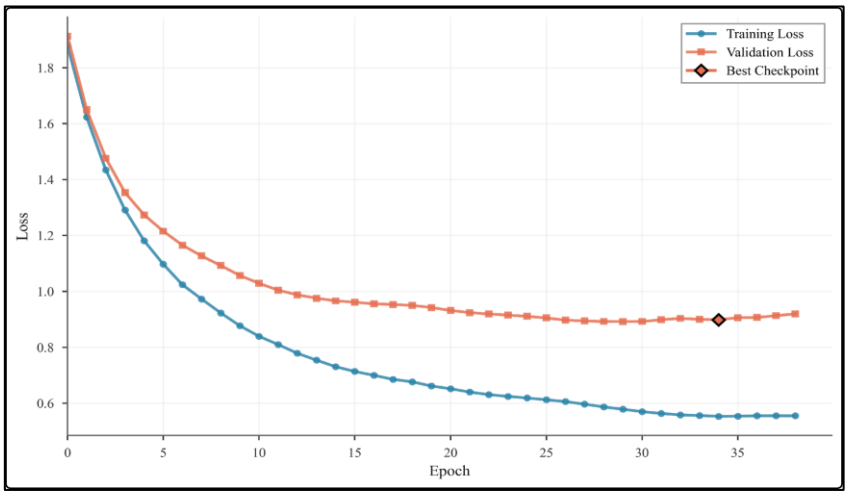

Figure 7. Training vs Validation Loss ASR Model

Figure 7 shows the training and validation loss curves across 38 epochs of the Conformer-based ASR model. The convergence of the curves happened quickly, with the training loss dropping from about 1.90 to 0.80 and the validation loss from 1.00 to 0.80 within the first 10 epochs. After this, the loss plateaus, with the training loss continuing to drop to 0.55 and the validation loss leveling off at about 0.90 past the 15-epoch mark. The best checkpoint was at epoch 34, and the gap between the curves is narrow from the very beginning, with the model able to generalize well with a minor overfitting effect.

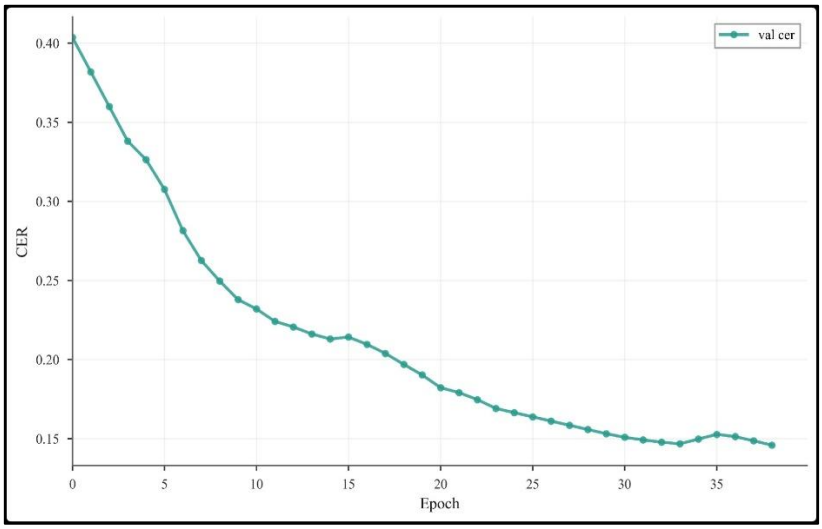

Figure 8. Validation Character Error Rate ASR Model

Figure 8: CER variation on the validation set as a function of epoch (38). The CER dropped from 0.40 to below 0.25 after 10 epochs and continued to decrease gradually until reaching approximately 0.146 at epoch 34.

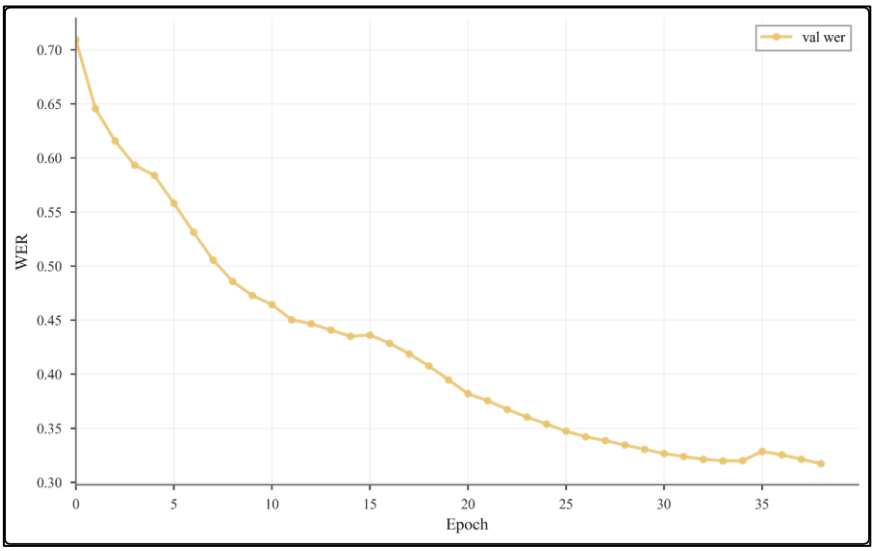

Figure 9. Validation Word Error Rate ASR Model

Figure 9 shows the change in the validation Word Error Rate over 38 epochs. It went from 0.72 to below 0.5 within the first 10 epochs, then gradually decreased until it was about 0.322 by epoch 34.

4.1.1 Ablation Study

An integrated system was set up by combining the speech transcription component with the emotion classification component. The goal is to transcribe spoken utterances accurately and estimate the emotional state of the speaker for better context-sensitive interaction. Transcription utilizes a large multilingual speech corpus. It covers both short- and long-range temporal dependencies, achieving, for a Conformer architecture, an error rate of 0.322 words and 0.146 characters. A number of emotional speech datasets have been collected for emotion recognition. After extracting various acoustic features under noisy conditions, the XGBoost model achieved an accuracy in emotion detection of 86.58%. These results illustrate the possibility of integrating speech transcription with emotion recognition and provide a basis for more human-like, empathic, and adaptable voice systems.

Table 1. Ablation of Data Augmentation Techniques on WER

Augmentation Type	Mean WER	Std. Dev. (over 5 seeds)	p-value
No augmentation	0.389	0.031	-
Time-mask only	0.379	0.033	0.18

Time-stretch only	0.385	0.036	0.42
Noise-mixing only	0.374	0.034	0.11
SpecAugment (time+freq)	0.362	0.029	0.04

4.2 Speech Emotion Recognition

The SER model based on XGBoost has strong performance on angry, neutral, happy, and surprise. Misclassifications mainly occur between the classes of fear and sad, and between happy and surprise. The overall accuracy is 0.8658, with a macro F1-score of 0.84. In weighted metrics, it achieved an F1-score of 0.866, precision of 0.87, and recall of 0.86.

These results suggest reliable emotion discrimination across classes with consistent recognition of major emotional states. The accuracy of 0.8658 reflects the model's ability to correctly classify a large proportion of utterances, while the macro F1-score of 0.84 shows balanced performance across all emotion categories. Emotions such as angry, happy, neutral, and sad are recognized with high confidence; this has mainly resulted in confusion between the fear-sad, neutral-sad, and disgust-sad pairs due to overlapping acoustic patterns and pitch contours. Strong results demonstrate that XGBoost effectively models the nonlinear dependencies among the acoustic features, which include MFCCs, pitch, and energy. Its framework of gradient boosting optimizes weak learners in a sequence to allow fine-grained separation between the classes of emotions. Additionally, the model efficiently handles class imbalance using sample weighting and performs well without relying on heavy GPU computation, making it suitable for small to medium-sized datasets.

Performance evaluation metrics were calculated using standard formulas:

$$Precision = (TP/TP + FP) \tag{1}$$

$$Recall = (TP/TP + FN) \tag{2}$$

$$\text{F1 Score} = (2 * Precision * Recall)/(Precision + Recall) \tag{3}$$

where TP denotes true positives, FP false positives, and FN false negatives.

Table 2. Performance Evaluation Metrics

Metrics	Value
Macro F1 Score	0.84
Weighted Precision	0.86
Weighted Recall	0.87
Weighted F1 Score	0.866

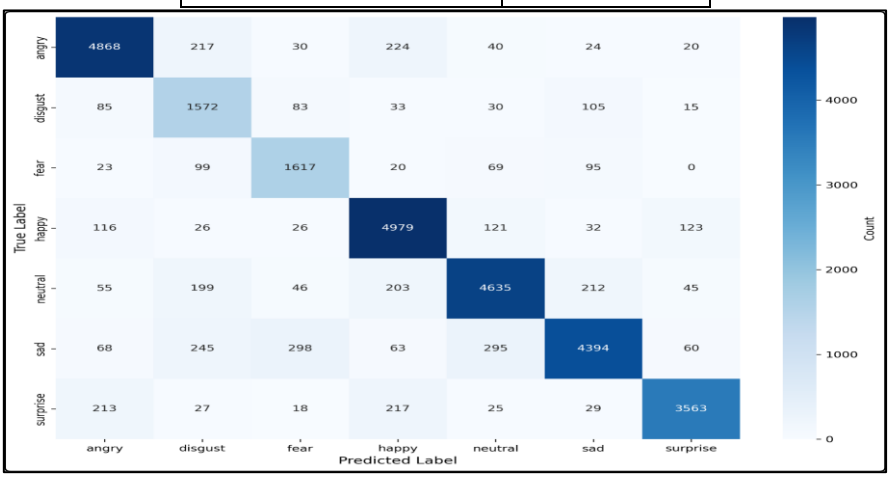

Figure 10. Confusion Matrix for SER Model

Figure 10 illustrates the confusion matrix for the emotion classification model on the test set, demonstrating strong diagonal dominance, indicating successful classification across all seven emotion categories. Angry, happy, neutral, and sad show the highest correct predictions, while notable confusion patterns are revealed between acoustically similar emotions, particularly among fear-sad, neutral-sad, and disgust-sad pairs, consistent with the overlapping acoustic characteristics of low-arousal negative emotions.

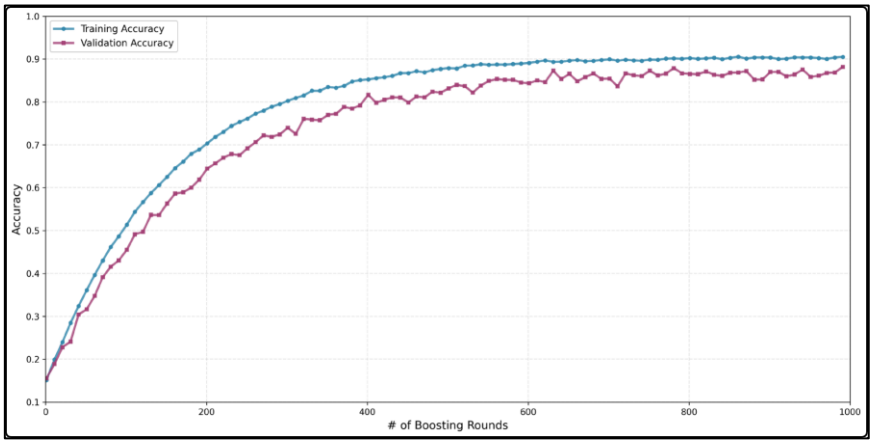

Figure 11. Training vs Validation Accuracy SER Model

Figure 11 illustrates the training and validation accuracy curves for a gradient boosting model over 1000 boosting rounds. Both curves show rapid improvement in the initial 200 rounds, with training accuracy rising from approximately 0.15 to 0.70 and validation accuracy following a similar trajectory. After round 400, the curves plateau, with training accuracy reaching 0.906 and validation accuracy stabilizing at 0.8658, indicating good model convergence with minimal overfitting as evidenced by the small gap between the two curves.

Table 3. Comparison of XGBoost with Conventional SER Models

Model	Accuracy
SVM	0.8444
Attentive CNN	0.6385
RNN with attention	0.635
XGBoost (proposed)	0.8658

Table 3 presents a comparison with other SER approaches reported in prior studies. Attentive CNN and RNN based models achieved accuracies of 0.6171 (Neumann and Vu [17]) and 0.635 (Mirsamadi, Barsoum, and Zhang [18]) respectively, while the SVM model reached about 0.8444(Zeng et al. [19]). XGBoost attains an accuracy of 0.8658 which outperforms these conventional methods under comparable experimental conditions.

5. Conclusion and Future Scope

The proposed framework includes an integrated approach comprising a Conformer-based ASR and an XGBoost-based SER system. Such a method will demonstrate how the linguistic and emotional features of speech can be jointly modeled to support context-aware human-centered interaction. The present work hence merges models of both into one unified system that will understand not only the spoken words but also their emotional context. This way, it will enrich education, customer service applications, and assistive communication with a never-before capability of machines acting with appropriate empathy and situational awareness. Thus, it forms a base for deeper research on speech multimodal systems that connect human emotion with machine understanding. Although promising, some improvements can be made to the proposed framework of Automatic Speech Recognition and Speech Emotion Recognition. First, larger and more diverse multilingual datasets will make the ASR system more robust against accents, noise, and speaking style variations. For SER,

incorporating multimodal cues like facial expressions, gestures, or physiological signals would help in disambiguating emotion detection when speech alone is uncertain.

References

[1] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is All You Need." Advances in neural information processing systems 30 (2017).

[2] Amodei, Dario, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper et al. "Deep speech 2: End-to-End Speech Recognition in English and Mandarin." In International conference on machine learning, PMLR, 2016, 173-182.

[3] Aouani, Hadhami, and Yassine Ben Ayed. "Speech Emotion Recognition with Deep Learning." Procedia Computer Science 176 (2020): 251-260.

[4] Heafield, Kenneth. "KenLM: Faster and Smaller Language Model Queries." In Proceedings of the sixth workshop on statistical machine translation, 2011,187-197.

[5] Kim, Tae-Wan, and Keun-Chang Kwak. "Speech Emotion Recognition Using Deep Learning Transfer Models and Explainable Techniques." Applied Sciences 14, no. 4 (2024): 1553.

[6] Kumar, Tapesh, Mehul Mahrishi, and Sarfaraz Nawaz. "A Review of Speech Sentiment Analysis Using Machine Learning." Proceedings of Trends in Electronics and Health Informatics: TEHI 2021 (2022): 21-28.

[7] Sahu, Gaurav. "Multimodal Speech Emotion Recognition and Ambiguity Resolution." arXiv preprint arXiv:1904.06022 (2019).

[8] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to Sequence Learning with Neural Networks." Advances in neural information processing systems 27 (2014).

[9] Vijayakumar, K. P., Hemant Singh, and Animesh Mohanty. "Real-Time Speech-To-Text/Text-To-Speech Converter with Automatic Text Summarizer using Natural Language Generation and Abstract Meaning Representation." Meaning Representation (2020).

[10] Ardila, Rosana, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. "Common Voice: A Massively-Multilingual Speech Corpus." In Proceedings of the twelfth language resources and evaluation conference, 2020, 4218-4222.

[11] Gulati, Anmol, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han et al. "Conformer: Convolution-Augmented Transformer for Speech Recognition." arXiv preprint arXiv:2005.08100 (2020).

[12] Livingstone, Steven R., and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English." PloS one 13, no. 5 (2018): e0196391.

[13] Cao, Houwei, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. "Crema-d: Crowd-Sourced Emotional Multimodal Actors Dataset." IEEE transactions on affective computing 5, no. 4 (2014): 377-390.

[14] Pichora-Fuller, M. Kathleen, and Kate Dupuis. "Toronto Emotional Speech Set (TESS)." Scholars Portal Dataverse 1 (2020): 2020.

[15] Jackson, Philip, and SJUoSG Haq. "Surrey Audio-Visual Expressed Emotion (Savee) Database." University of Surrey: Guildford, UK (2014).

[16] Zhou, Kun, Berrak Sisman, Rui Liu, and Haizhou Li. "Emotional voice conversion: Theory, Databases and Esd." Speech Communication 137 (2022): 1-18.

[17] Neumann, Michael, and Ngoc Thang 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).

[18] Mirsamadi, Seyedmahdad, Emad Barsoum, and Cha Zhang. "Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention." In 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP), IEEE, 2017, 2227-2231.

[19] Zeng, Xiaoping, Li Dong, Guanghui Chen, and Qi Dong. "Multi-Feature Fusion Speech Emotion Recognition Based on SVM." In 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC), IEEE, 2020, 77-80.