Comparative Analysis of Speech Emotion Recognition System Using MLP, SVM, and CNN Algorithms

*Omodunbi B.A., Awoyemi T.A., Okomba S.N. & Esan A.O.
Department of Computer Engineering, Federal University, Oye-Ekiti
*Corresponding author: awoyemitaiwoakinol@gmail.com

Received: 21/09/2025 **Revised:** 19/10/2025 **Accepted:** 4/11/2025

Emotion recognition from speech plays a crucial role in enhancing human–computer interaction by enabling systems to interpret and respond to users' emotional states. This study develops and evaluates a Speech Emotion Recognition (SER) system using three machine learning techniques; Support Vector Machines (SVM), Multilayer Perceptron (MLP), and Convolutional Neural Networks (CNNs). The system is trained and tested on the RAVDESS dataset, which contains 1,440 professionally recorded audio samples representing a wide range of emotions. Our approach involves careful preprocessing of the audio signals, extraction of key acoustic features, and comparative performance evaluation of the three models using standard metrics. Results show that each model exhibits unique strengths and limitations, with CNNs achieving the most robust feature learning and generalization. The study underscores the importance of diverse feature representation for accurate emotion classification and provides insight into how different model architectures handle emotional nuances in speech. Identified challenges such as dataset diversity, feature selection, and computational complexity are discussed, along with recommendations for future research to improve SER systems' real-world adaptability. This work contributes to ongoing efforts toward developing emotionally aware technologies that can enhance natural human–machine communication.
Keywords: Speech, speech recognition, pre-processing, evaluation techniques, communication

Introduction

Effective communication relies on the ability to express oneself, and humans employ a combination of body language, hand gestures, and vocal nuances to communicate effectively (Chen, 2020). These elements collectively convey emotions and sentiments (Yu, 2019). While verbal communication may vary across different languages worldwide, the nonverbal aspect of communication, which expresses feelings, is largely universal. Progress in emotion detection technology has made a significant impact on fields such as psychology, psychiatry, and neuroscience research (Abeer, 2019). Within cognitive sciences, human interaction plays a central role, where subjects are often presented with a range of questions or scenarios, and their responses form the basis for various inferences. A challenge arises, however, with individuals such as introverts, who may be hesitant to express themselves openly. In such cases, substituting traditional methods with computer-based systems provides an avenue for more objective and comprehensive emotional assessment (Wu *et al*., 2019). Beyond research, the practical applications of speech emotion recognition extend to customer service, healthcare, security, and human–computer interaction, making it a critical area of study. By allowing systems to identify and respond to human emotions, speech emotion recognition enhances how machines interact with people, enabling more natural and empathetic communication in both professional and personal environments.

Despite these advancements, speech emotion recognition still faces limitations that restrict its effectiveness in real-world use. Current approaches often struggle with background noise, speaker variability, and cross-cultural differences in vocal expression, leading to reduced accuracy and poor generalization outside controlled environments. Although the present study uses the RAVDESS dataset, which contains high-quality studio recordings by professional actors, this controlled data provides a necessary foundation for consistent evaluation of models before extending future work to real-world conditions. To address these challenges, the study focuses on developing and evaluating a speech-based emotion recognition system that combines three techniques; Support Vector Machine, Multilayer Perceptron, and Convolutional Neural Network. The specific research questions guiding this work are: which of these algorithms performs best in identifying emotional cues from speech, which features contribute most to classification accuracy, and how do the strengths of each model differ in terms of learning and generalization. The novelty of this study lies in its systematic comparison of traditional and deep learning methods under uniform pre-processing and evaluation settings, providing a deeper understanding of their behaviour and suitability for emotion classification. Ultimately, the goal is to improve recognition accuracy and contribute to the broader vision of emotion-aware technologies that can enhance human–computer

interaction, enrich psychological research, and support more effective communication across various domains.

Related Works

Several researchers have explored the use of deep learning in emotion-driven systems, extending beyond speech-based recognition into multimodal applications. For instance, Omodunbi *et al*. (2023) developed a facial emotional-based song suggestion system using a convolutional neural network, which automatically recommends music according to the listener's detected mood. Their model, trained on the FER2013 dataset, demonstrated how CNNs can effectively capture and classify subtle emotional expressions from facial cues. While this work focuses on music recommendation, its methodology further validates the strength of CNN architectures in emotion recognition tasks, aligning with findings in speech emotion recognition and reinforcing the model's adaptability across domains.

In the paper titled "Survey of Technical Progress in Speech Recognition by Machine over Few Years of Research" by Okomba *et al*. (2019), a comprehensive overview of research efforts in Automatic Speech Recognition (ASR) conducted over the past decade is presented. The paper delves into key themes and advancements achieved during this period, providing insights into the evolving technological landscape and recognizing the significant progress made in the field of speech communication. Despite contextual variations, diverse environmental conditions, speaker diversity, and lower-quality audio, attaining high accuracy in automatic speech recognition has remained a pivotal research challenge throughout the course of research and development. Designing a speech recognition system necessitates careful consideration of various elements, including defining different speech categories, selecting methods for speech representation, employing specific techniques, choosing suitable databases, and utilizing methods for performance evaluation.

In the paper 'Two-way Feature Extraction for Speech Emotion Recognition using Deep Learning' by Abeer (2019), a novel approach to effective speech emotion recognition is presented. Initially, a two-way feature extraction method is introduced, utilizing super convergence to derive two distinct sets of potential features from the speech data. The first set of features undergoes Principal Component Analysis (PCA) to obtain the initial feature set. Subsequently, a Deep Neural Network (DNN) incorporating dense and dropout layers is employed. In the second approach, mel-spectrogram images are generated from audio files, and these 2D images serve as input to a pre-trained VGG-16 model. The study conducts extensive experiments and conducts a thorough comparative analysis of both features extraction methods, employing multiple algorithms and evaluating on two different datasets.

Another study conducted by Wu *et al*. (2020), focused on analysing baby crying using MFCC and LFCC in different classification methods. The research involved pre-processing, feature extraction, and classification stages. They compared the performance of KNN classification, Vector Quantization, and Simple Neural Network using both MFCC and LFCC. The results showed that KNN classification with LFCC achieved higher accuracy than using MFCC. The study included various categories of baby sounds and noise additions, and LFCC demonstrated an average accuracy of 91.58%, while MFCC had an average accuracy of 82.14%.

In a comparative study by Schuller *et al*. (2021), different feature extraction techniques for automatic speech recognition were analysed using a multi-criteria process. The objective was to identify discriminative and robust features in the acoustic data. The study employed a weighted score method (WSM) to compare the extraction techniques based on various criteria. The results showed that MFCC was effective for isolated word speech but suffered from deficiencies in noisy environments. Other techniques like PCA, DWT, and RASTS-PLP were effective in minimizing noise. The comparison highlighted the complementary nature of the extraction techniques, suggesting the need for hybrid approaches.

Lastly, Yu *et al* (2021) proposed a method for extracting speech features using MFCC in an automatic speech recognition system. They applied Mel-frequency cepstral coefficient feature extraction and Mel-frequency filtration processes to enhance speech and reduce background noise. The proposed method improved the recognition process and minimized distortion in the voice. Additionally, they explored delta derivatives of MFCC and observed that applying the MFCC technique reduced distortion in second-order delta coefficients, resulting in better performance. They recommended the use of spectrogram windowing with a hybrid approach of deep learning algorithms to further enhance the system's performance.

Research Methodology

This section outlines the proposed methodology, the emotion database, the classifier model, and the workflow for extracting emotions from the dataset.

Dataset

The dataset used in this study is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). It contains 1,440 audio recordings in .wav format and was created specifically for emotion recognition research. The recordings were performed

by 24 professional actors, each delivering speech that captures a wide range of emotions. These emotions include calm, happy, sad, angry, fearful, surprised, disgusted, and neutral states. For most emotions, two levels of intensity were recorded, normal and strong, while the neutral expression appears only in its baseline form. To prepare the dataset for training and evaluation, the data was split into two parts: 70 percent was used for training the models, while the remaining 30 percent was set aside for testing. This ensured that the models were trained on a large portion of the data while still being tested on unseen samples to properly assess their performance.

Feature extraction

The dataset used for this study contains different types of recordings, including speech, songs, and video clips with sound. Since raw audio cannot be directly applied for emotion recognition, pre-processing was first carried out to improve the quality of the input. This involved resampling all audio signals to a consistent frequency, normalizing amplitude levels to reduce volume variability, and trimming silent segments to focus on active speech. Noise reduction techniques were also applied to minimize background interference, thereby ensuring that the extracted features represent only the most relevant aspects of the signal. For feature extraction, Chroma was selected as the primary descriptor of emotional content in speech. Chroma features capture the pitch class profile of an audio signal, breaking the sound into 12 distinct pitch classes and showing how they evolve over time. This makes them particularly effective in identifying changes in intonation and melody, which are strong indicators of emotional state. Chroma consists of two major components: the Chroma Vector, which quantifies the strength of each pitch class at a given time, and the Chroma Deviation, which reflects how much the pitch distribution fluctuates. Although other features such as Mel Frequency Cepstral Coefficients (MFCC), Linear Frequency Power Coefficients (LFPC), spectral contrast, and zero crossing rate are widely recognized in emotion recognition research for their robustness in capturing timbral and spectral information, this study intentionally focuses on Chroma alone to evaluate its standalone effectiveness in identifying emotional cues through pitch and tonal variation. The rationale behind this choice lies in Chroma's strong correlation with musical and prosodic features that directly mirror emotional expression in human speech, making it a valuable feature for isolating the role of tonality independent of other spectral characteristics. This narrowed focus is acknowledged as a limitation, as incorporating additional features like MFCC or LFPC could potentially enhance recognition accuracy. The extracted Chroma features were then fed into the training pipeline, where configurations such as batch size, number of epochs, and learning rate were fine tuned to achieve optimal model performance.

Classification

Two machine learning algorithms were used for the classification of the various emotions.

MLP classifier

The system is built using three main layers: the input layer, the hidden layer, and the output layer. The input layer serves as the entry point, where the extracted Chroma features from the audio signals are received and prepared for processing. This information is then passed to the hidden layers, which are responsible for learning complex, non-linear relationships between the features that are not immediately visible in the raw data. In this study, the Multilayer Perceptron (MLP) architecture consists of two hidden layers with 128 and 64 neurons respectively, allowing the model to progressively refine its internal representation of the emotional patterns present in speech. Each hidden layer employs the Rectified Linear Unit (ReLU) activation function to introduce non-linearity and prevent vanishing gradients during training. The output layer uses the Softmax activation function to generate probability distributions across the emotion categories; happy, sad, angry, and neutral, enabling the system to select the most likely emotion class. The hidden layers are termed "hidden" because they are not directly connected to the outside world; instead, they interact only with the layers before and after them, enabling internal learning of abstract emotional features. This configuration ensures a balance between model complexity and computational efficiency, supporting stable convergence and reproducibility of results.

SVM classifiers

Support Vector Machines (SVMs) are models that classify data by drawing hyperplanes in a high-dimensional feature space, serving as decision boundaries that separate data points into distinct groups based on shared patterns. To determine the optimal boundary, SVMs rely on algorithms grounded in statistical learning theory, ensuring maximum separation between classes. Since SVMs operate under supervised learning, they require labeled training data. After extracting features from the preprocessed audio signals, the SVM model was trained using a radial basis function (RBF) kernel, which effectively captures nonlinear relationships within the data. The regularization parameter and kernel coefficient were fine-tuned to balance model bias
and variance, leading to improved classification accuracy. The final configuration enabled the SVM to accurately map the extracted features to their corresponding emotional categories.

Figure1: Workflow of SER using SVM and MLP Classifier

Workflow

The workflow in Figure 1 outlines the structured process used to develop the speech emotion recognition system. It begins with data collection to obtain quality audio samples, followed by feature extraction, where raw audio is transformed into meaningful representations such as Chroma features. These features are then used in the training phase, allowing the models to learn emotional patterns. Classification is performed using Support Vector Machines (SVM) and Multilayer Perceptron (MLP), and finally, the system's performance is evaluated using standard metrics. This workflow was chosen for its simplicity, logical flow, and ability to ensure consistent and reliable emotion recognition results.

System implementation

Load dataset

The implementation began with loading the dataset into the Python environment using pandas and librosa libraries. Each audio file was programmatically accessed, and its corresponding emotion label was extracted from the filename structure. The data was then organized into arrays for input features and target outputs. This setup ensured that each emotion class was properly indexed and balanced before feature extraction.

Feature extraction

Feature extraction was carried out using the librosa library, which converted each audio signal into chroma representations. The process included normalizing amplitude, trimming silence, and resampling all audio files to a consistent frequency. The chroma features were computed with a fixed frame size and hop length, and the resulting feature vectors were stored as NumPy arrays for model training. This automated pre-processing pipeline streamlined the conversion of raw audio into usable input data

Classification

The Two main models were implemented for emotion classification: Support Vector Machine (SVM) and Multilayer Perceptron (MLP). Both models were developed and trained using the scikit-learn library. The SVM used a radial basis function kernel, while the MLP employed the ReLU activation function with a single hidden layer. Regularization and kernel parameters were fine-tuned experimentally to maximize classification accuracy. The models were trained and tested on an 80:20 split, and their performance metrics, including accuracy and F1-score, were compared to assess effectiveness.

Evaluation metrics

These are different ways of training and testing the performance of the models for the given dataset. Selection of appropriate evaluation metrics is important for proper understanding of a model. Accuracy, precision, Recall and F1 score are the metrics used for evaluation of the speech emotion recognition model.

Accuracy

The most fundamental evaluation criterion used to gauge a model's performance is accuracy as expressed in Equation 1. It is determined as the total number of forecasts divided by the number of accurate ones. It assesses a model's capacity for accurately classifying all pixels, whether they are positive or negative.

$$accurary = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \quad 1$$

Precision

For problems with imbalanced binary classification, two assessment measures are frequently used: precision and recall. Precision is defined as the ratio of true positive predictions to all positive predictions as shown in Equation 2.

$$precision = \frac{Tp}{Tp+Fp} \quad 2$$

Recall

On the other hand, recall is the percentage of accurate positive predictions across all emotions in the datasets expressed in Equation 2. While recall and precision are connected, they focus on separate facets of model performance. A model with high accuracy makes accurate positive predictions, whereas a model with high recall correctly identifies the majority of speeches with correct output

$$recall = \frac{Tp}{Tp+Fn} \quad 3$$

F1-Score

The F1-score amalgamates recall and precision into a unified metric. By computing the harmonic mean of recall and precision, it strikes a balance between the two measures. A model with a high F1-score not only adeptly identifies the majority of emotions classifications, but also ensures accurate positive predictions.

$$F1 = \mathbf{2} * \frac{1}{1/precision + 1/recall}$$

Results and Discussion

This paper presents the development of different models, followed by a results analysis carried out using standard evaluation metrics. The outcomes were compared with respect to the feature extraction techniques applied, highlighting how each approach influenced model performance. In addition, the accuracy, Recall, F1-Score of the models when tested is illustrated in Table 1 and 2 respectively.

Table 1: Accuracy of the models

MODELS	ACCURACY (%)
Support Vector Machine	80
Logistic Regression	86
XGBoost	85
MLP	88
CNN	90

Table 2: Different Evaluation Metrics of the models

	SVM	MLP	XGBOOST	CNN	LOGISTIC REGRESSION
Precision	0.80	0.88	0.85	0.90	0.86
Recall	0.81	0.88	0.85	0.91	0.86
F1-Score	0.80	0.88	0.85	0.90	0.86

The results of this study show how the different machine learning models performed in recognizing emotions from speech. The comparison included Support Vector Machine, Multilayer Perceptron, Logistic Regression, XGBoost, and Convolutional Neural Networks. Among these, the convolutional model achieved the highest accuracy of 90 percent, confirming its strength in identifying complex patterns within audio signals. The Multilayer Perceptron followed with 88 percent, while Logistic Regression and XGBoost achieved 86 and 85 percent respectively. The Support Vector Machine recorded the lowest accuracy of 80 percent, indicating difficulty in adapting to nonlinear variations in the dataset. Analysis of the confusion matrices showed that happiness and neutral emotions were classified with the highest accuracy, while fear and disgust were the most frequently misclassified due to their overlapping acoustic traits. The per-emotion evaluation demonstrated that the convolutional model maintained a stable performance across all emotion categories, while the Support Vector Machine and Logistic Regression showed reduced precision when separating sadness from calm expressions.

The training analysis showed that the convolutional model required about 420 seconds per training cycle, the Multilayer Perceptron 270 seconds, and the Support Vector Machine 130 seconds. Once trained, the convolutional model made faster predictions, which makes it suitable for real-time tasks. Its superior accuracy comes from its ability to learn spatial relationships in the speech spectrogram, allowing it to capture changes in tone, rhythm, and intensity that other models could not interpret effectively. These findings confirm that while traditional algorithms can classify emotions in speech, models based on layered neural networks achieve greater recognition consistency and practical efficiency, making them more effective for use in emotion-aware systems.

Conclusion

This study examined the use of different machine learning models for recognizing emotions in speech, with Support Vector Machine and Multilayer Perceptron as the main focus. Additional models such as Logistic Regression, XGBoost, and Convolutional Neural Networks were also evaluated to provide a broader comparison. The results showed that the convolutional model achieved the highest accuracy at 90 percent, while the Multilayer Perceptron followed with 88 percent. Although these figures suggest that neural network methods capture emotional cues more effectively than traditional algorithms, the margin of difference between MLP and SVM was modest and

may not indicate a statistically significant improvement. The findings show that while deep learning models have clear advantages in representation learning, traditional methods like SVM still hold potential when properly tuned and optimized for feature scaling and class balance. These results underline the importance of both model architecture and data preparation in achieving reliable emotion recognition from speech.

Future research should build on these outcomes by targeting specific areas where models showed weakness. The confusion analysis revealed that emotions such as fear and disgust were often misclassified, indicating the need for targeted data augmentation to strengthen recognition in those categories. Expanding the dataset with more diverse voices and natural speech conditions would also support better generalization. Further experimentation with adaptive learning rates, advanced neural layers, and hybrid architectures could enhance both accuracy and robustness. A focused approach that combines data diversity with architecture optimization would contribute to more dependable and practical emotion-aware systems.

References

Abeer, S., Raza, H., & Qamar, U. (2019). Speech emotion recognition using deep convolutional neural networks. *Procedia Computer Science, 152*, 407–414. https://doi.org/10.1016/j.procs.2019.05.055

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. In *Proceedings of Interspeech* (pp. 1517–1520).

Cao, R., Verma, A., & Nenkova, H. (2019). Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech. *Computer Speech & Language, 28*(1), 186–202. https://doi.org/10.1016/j.csl.2013.06.002

Chakraborty, I., & Sharma, G. (2018). Speech emotion recognition: A comparative analysis of datasets and features. *Journal of King Saud University - Computer and Information Sciences, 30*(3), 1–11. https://doi.org/10.1016/j.jksuci.2016.12.004

Chen, X., Mao, Y., Xue, L., & Cheng, L. L. (2020). Speech emotion recognition: Features and classification models. *Digital Signal Processing, 22*(6), 1154–1160. https://doi.org/10.1016/j.dsp.2012.09.006

Cheng, P., Zhang, G., Schuller, B., & Zafeiriou, S. (2019). End-to-end speech emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing, 11*(8), 1301–1307. https://doi.org/10.1109/JSTSP.2019.2952335

Dai, W., Han, D., Dai, Y., & Xu, D. (2020). Emotion recognition and affective computing on vocal social media. *Information & Management, 57*(5), 103223. https://doi.org/10.1016/j.im.2019.103223

Loan, R. (2018). Emotional speech recognition using deep neural networks. *Cognitive Computation, 10*(3), 448–460. https://doi.org/10.1007/s12559-018-9554-2

Loan, Y., Cao, W., Zhang, Z., & Wang, D. (2019). A hybrid model based on CNN and BiLSTM for speech emotion recognition. *IEEE Access, 9*, 27098–27107. https://doi.org/10.1109/ACCESS.2019.2897260

Nwe, S. W., Foo, L. C., & De Silva, T. L. (2020). Speech emotion recognition using hidden Markov models. *Speech Communication, 41*(4), 603–623. https://doi.org/10.1016/S0167-6393(03)00099-2

Okomba, S., Adegboye, M., & Candidus, O. (2019). Survey of technical progress in speech recognition over recent years. *Computer Engineering Journal, 15*(2), 407–414.

Omodunbi, B. A., Soladoye, A. A., Olaniyan, O. M., Salami, A. I., & Olagunju, A. I. (2023). Facial emotion-based song suggestion system using convolutional neural network. *International Journal of Advanced Computer Science and Applications, 14*(5), 1–10. https://doi.org/10.14569/IJACSA.2023.0140501

Schuller, B., Steidl, S., & Batliner, A. (2009). The INTERSPEECH 2009 emotion challenge. In *Proceedings of Interspeech* (pp. 312–315).

Wu, C.-H., & Liang, W.-B. (2019). Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction* (pp. 1–6). Tainan, Taiwan.

Wu, S., Falk, T. H., & Chan, W.-Y. (2020). Automatic speech emotion recognition using modulation spectral features. *Speech Communication, 53*(5), 768–785. https://doi.org/10.1016/j.specom.2010.02.009

Yu, F., Zhang, L., & Li, H. (2021). Emotion detection from speech to enrich multimedia content. In *Proceedings of the Pacific-Rim Conference on Multimedia (PCM)* (pp. 1–10). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-12344-0_8

Yule, Y. H., & Hsu, W. H. (2019). An efficient speech emotion recognition system using a hybrid model. *IEEE Transactions on Audio, Speech, and Language Processing, 21*(12), 2570–2580. https://doi.org/10.1109/TASL.2013.2278904

Zhang, Y., Zhang, D., Wang, S., & Liu, Z. (2019). Speech emotion recognition using 1D convolutional neural networks. In *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). https://doi.org/10.1109/IJCNN.2019.8852014