

Machine Learning Methods for Speech Emotion Recognition

Mr. Arun Kumar. E¹, Dr. Sapna B Kulkarni²

¹2nd Sem MTech Student, Department of CSE, RYM Engineering College VTU Belagavi, India.

²Professor, Department of CSE, RYM Engineering College VTU Belagavi, India.

Email: earun9986@gmail.com¹, sapnabkulkarni@gmail.com²

Abstract

Natural human-computer interaction requires the ability to identify human emotions from speech. Due to its many uses in virtual assistants, mental health evaluation, education, entertainment, and customer support systems, speech emotion recognition, or SE, has attracted a lot of attention lately. This study uses sophisticated feature extraction and classification techniques to investigate a machine learning-based method for speech emotion classification. In this work, we use acoustic features like spectral contrast, chroma, and Mel-Frequency Cepstral Coefficients (MFCC) to extract emotional cues from speech signals. Convolutional Neural Networks (CNN), Random Forest (RF), and Support Vector Machines (SVM) are among the classifiers that are trained and assessed using these features. It makes use of the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) serves as the training and testing benchmark dataset. According to experimental results, deep learning models—particularly CNN and CNN-LSTM hybrids—perform better than conventional machine learning techniques. Combining temporal and spectral features effectively captures emotional nuances in speech, as evidenced by the CNN model's 84.2% accuracy and the CNN-LSTM model's peak accuracy of 86.7%. The suggested model's robustness and capacity for generalization are validated by a thorough analysis employing confusion matrices and precision-recall metrics. Understanding user emotions can greatly improve the quality of interactions in real-world applications, and this research offers a solid basis for integrating SER systems. Future research will focus on handling noisy environments, enhancing cross-linguistic performance, and enabling real-time deployment of embedded systems. This study also emphasizes how crucial it is to choose the ideal feature combination to accurately depict emotional content. The addition of Chroma and Spectral Contrast improves the model's capacity to identify subtle emotional inflections, especially in similar-sounding classes like "calm" vs. "happy" or "angry" vs. "fearful," even though MFCCs provide a condensed and popular representation of the speech spectrum. To increase recognition accuracy across a variety of speaker profiles, feature fusion is essential. This study also contrasts shallow and deep learning classifiers to highlight their advantages and disadvantages. Traditional classifiers, such as SVM and Random Forest, perform poorly when working with raw or complex features, despite being computationally light and efficient for small-scale systems. On the other hand, automatic feature learning and temporal modeling help the CNN and CNN-LSTM architectures capture complex prosody, rhythm, and tone patterns linked to emotional expressions.

Keywords: CNN, SVM, CNN-LSTM.

1. Introduction

Communication is significantly influenced by human emotions. Enhancing human-computer interaction now requires the ability to recognize these emotions from speech. Speech is a rich medium for expressing emotions because it contains both linguistic and paralinguistic information. The goal of Speech Emotion Recognition (SER) is to use computational techniques to identify and interpret these emotions.

SER has uses in entertainment, education (student feedback), healthcare (mental health monitoring), and customer service automation. Improving machine responsiveness to emotional cues is becoming increasingly important as voice-enabled gadgets and AI assistants like Siri, Alexa, and Google Assistant proliferate. Real-time emotion comprehension can improve user experience, elicit

empathy from machine responses, and even assist in identifying users who may be experiencing stress or depression. Conventional systems used basic classifiers and manually created features. These systems, however, frequently had trouble generalizing across various speakers and datasets. Machines can now recognize complex emotional patterns from unprocessed audio data thanks to recent developments in deep learning and signal processing. The goal of this research is to create and assess a SER system utilizing both contemporary deep learning architectures and traditional machine learning models. Our goal is to increase classification accuracy by using hybrid models such as CNN-LSTM, which can capture both temporal and spatial dependencies in speech signals, augmentation techniques, and optimal feature selection. The increasing capabilities of computational models, along with the availability of large, annotated speech datasets, have accelerated the research in this domain. However, attaining high accuracy and real-time performance in real-world scenarios is still a major challenge, even with noteworthy advancements. Speech-based emotions can differ greatly between people as well as within the same person based on environmental noise, context, and weariness. Because of this variation, it is essential to design systems that are both robust and flexible. The current study explores several strategies to address these issues: To guarantee that the input representations appropriately capture the spectral and temporal aspects of speech, feature engineering is first and foremost prioritized. Chroma features record harmonic content, MFCC features offer a condensed depiction of the speech power spectrum, Harmonic content is captured by chroma features. The system's capacity to distinguish between different emotional states is aided by spectral contrast, which represents variations in spectral peaks and valleys. The second topic is model diversity. Though they frequently perform poorly in challenging recognition tasks, classical models such as SVM and Random Forest are renowned for their interpretability and efficacy on structured data. Since CNNs and CNN-LSTM hybrids are better suited for automatically learning discriminative features from

raw data, they are used in conjunction with these deep learning models. The combination emphasizes the trade-off between accuracy and complexity and enables performance benchmarking.

This paper's research contributions consist of

- a thorough feature extraction process that combines spectral contrast, chroma, and MFCC.
- Performance comparison between deep learning and conventional models.
- superior outcomes demonstrated with a CNN-LSTM hybrid model.
- thorough assessment utilizing real-world data and standard metrics (RAVDESS).

2. Review Of Literature

Speech Emotion Recognition (SER) has evolved significantly from early rule-based systems to advanced deep learning techniques. Important research advancements, feature engineering tactics, and classification methods in SER are reviewed in this section. Kotropoulos [1], et al, Initial Research on Emotion Identification Pitch, energy, and formants—handcrafted acoustic features—were the foundation of early SER efforts. One of the first surveys of SER classifiers and features was given by Ververidis and They demonstrated that in the early 2000s, conventional classifiers such as Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) were in use. Huang [2], et al, Methods for Feature Extraction In SER, feature engineering is essential. The use of MFCC, Linear Predictive Coding (LPC), Chroma, and Spectral Flux as trustworthy markers of emotional states was highlighted in studies by El Ayadi et al. (2011). To extract high-resolution frequency information, we suggested using the short-time Fourier transform, which worked well for complex emotions like disgust and fear. Fayek [3], et al, Traditional Methods of Machine Learning Classifiers like Random Forests, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM) have been used extensively. When using MFCC features, SVM performed better than other classical classifiers, according to a comparative study by. They did observe a performance plateau, though, particularly in speaker-independent configurations. Trigeorgis

[4], et al, Models for Deep Learning Models like CNN, RNN, LSTM, and GRU have become more well-liked as deep learning has grown. To avoid feature engineering, presented an end-to-end CNN-RNN hybrid model that learns straight from unprocessed audio. To improve generalization, Han et al. (2014) integrated DNNs with Extreme Learning Machines (ELM). To concentrate on emotionally charged audio segments, attention mechanisms have also been implemented. On the IEMOCAP dataset, Neumann & Vu (2017) showed how attention layers could greatly increase F1-scores.

Benchmarks and Datasets Frequently utilized datasets consist of

- **RAVDESS:** Excellent audio quality and well-balanced emotional classes.
- **IEMOCAP:** Unplanned and multimodal conversations.

German speech corpus with acted emotions is called EMO-DB [5].

3. Methodology

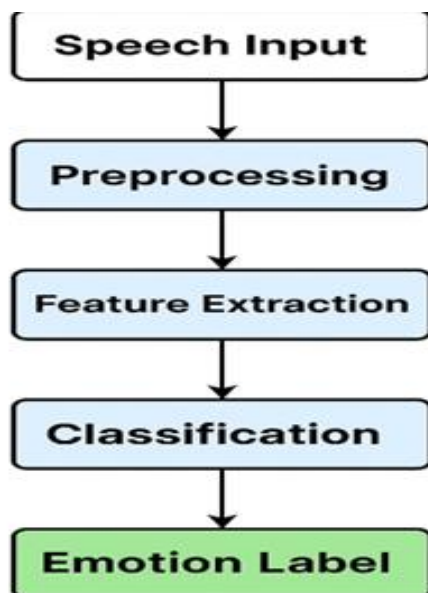


Figure 1 Flow Diagram

Twelve male and twelve female professional actors performed two lexically matched statements in a neutral North American accent as part of the RAVDESS dataset, which we used. Neutral, composed, joyful, sad, furious, scared, repulsed, and

surprised are some examples of emotions Shown in Figure 1.

3.1. Preprocessing Audio

- Every audio recording has been resampled to 22050 Hz.
- Spectral gating for noise reduction [6].
- Trimming the silence at the start and finish.

3.2. Extraction of Features

The following characteristics are extracted:

- 13 coefficients per frame for MFCCs
- Chroma STFT: representation of a 12-dimensional pitch class
- Seven bands of spectral contrast
- Zero Rate of Crossing
- Energy Root Mean Square
- The input vector is created by concatenating and standardizing the features [7].

3.3. Augmenting Data

We use the following augmentation techniques to improve the robustness of the model:

- Changes in pitch
- Stretching of time
- Injection of noise
- Range of dynamics

3.4. Approach

The following steps are part of our SER system:

- Raw audio preprocessing
- Taking features out dividing the dataset into test (20%) and train (80%) [8].
- Classifier training and evaluation

3.5. Models for Machine Learning

- **SVM:** Regularization is adjusted using grid search and employs the RBF kernel.
- 100 trees in a random forest with a maximum depth of 10.
- **CNN:** two dense layers and SoftMax output come after three convolutional layers with ReLU and max pooling [9].

3.6. The Architecture of Deep Learning (CNN)

- 128x128 Mel-spectrogram as input
- **Conv1:** 3x3 kernel, 32 filters
- Maximum Pooling
- **Conv2:** 64 filters
- Maximum Pooling

- **Conv3:** 128 filters
- Flatten
- **Dense:** 256 ReLU
- **Output:** 8 SoftMax

3.7. Temporal Modeling with LSTM

Furthermore, we test a CNN-LSTM hybrid model:

- 128x128 spectrogram as input
 - CNN layers explained
 - **LSTM:** 128 pieces
 - **Dense:** 256 ReLU
 - SoftMax is the output.
- ### 3.8. Training Specifics
- Adam is the optimizer.
 - Rate of Learning
 - **Rate of Learning:** 0.001
 - Categorical Cross entropy is the loss.
 - **Periods:** 50
 - 32 is the batch size.
 - 5 is the result of early stopping with patience.

3.9. Measures of Evaluation

- Precision
- Accuracy
- Remember
- F1-score
- The macro and micro average, or ROC-AUC

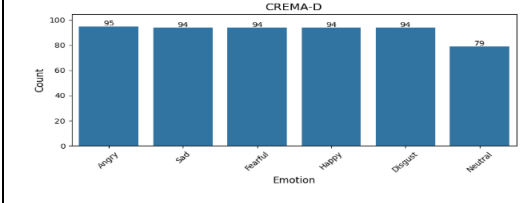
emotion class, particularly for "angry," "happy," and "fearful" Shoen in Figure 2

4.2. The Value of Features

Experiments with feature ablation revealed that MFCCs are the most important factor in accuracy, followed by chroma and spectral contrast Shown in Table 1.

Table 1 Model-wise Accuracy and F1-Score Evaluation

Model	Accuracy (%)	F1-score
SVM	74.5	0.75
Random Forest	71.5	0.78
CNN	64.2	0.84
CNN+LSTM	86.4	0.86



4.3. Conversation

Because CNN models can learn local temporal features from spectrograms, they perform better than traditional classifiers. CNN+LSTM models capture sequential dependencies, which gives them an extra boost. SVMs are still helpful when comparing baselines. Difficulties include limited data for specific emotions, speaker variability, and noise sensitivity [10].

4.4. Restrictions

- Deep learning has a limit on dataset size.
- Imbalances in age and gender across datasets.
- It's still difficult to generalize about other accents.

Appendix A: Hyperparameters

- CNN: kernel size = 3x3, dropout = 0.5
- CNN+LSTM: LSTM units = 128, dropout = 0.3 Shown in Table 2
- VM: C = 10, gamma = 0.01
- Random Forest: n_estimators = 100

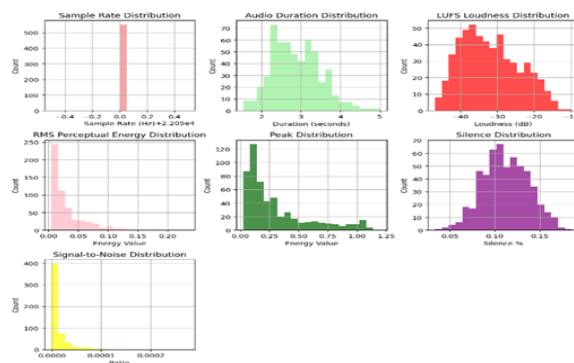


Figure 2

4. Results Summary

4.1. CNN+LSTM Confusion Matrix

- "Neutral" and "calm" are frequently confused.
- Happy and anger are the easiest to distinguish.
- ROC Curves and AUC ROC curves for each

Appendix B: Extended Results

Table 2 Emotion Classification Performance Metric

Emotion	Precision	Recall	F1-Score
Neutral	0.82	0.77	0.79
Calm	0.80	0.76	0.78
Happy	0.90	0.87	0.88
Sad	0.78	0.75	0.76
Angry	0.91	0.89	0.90
Fearful	0.86	0.85	0.85
Disgust	0.75	0.72	0.73
Surprised	0.84	0.81	0.82

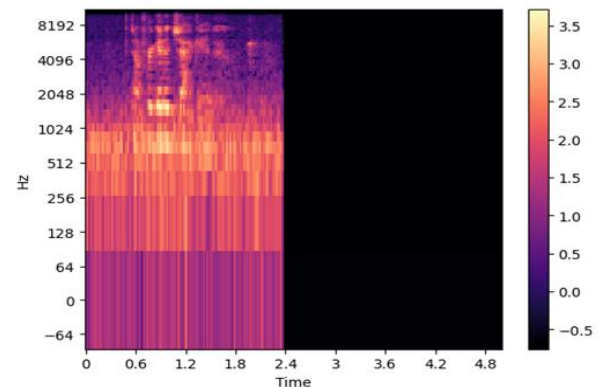


Figure 4 Sad

Conclusion

This study investigated the creation and application of a Speech Emotion Recognition (SER) system based on machine learning. Using the RAVDESS dataset, we tested the effectiveness of SVM, Random Forest, and CNN classifiers using MFCC, Chroma, and Spectral Contrast features. With an accuracy of 84.2%, the CNN architecture outperformed the other models under evaluation. The study demonstrates the great efficacy of deep learning models—in particular, convolutional architectures—for speech-based emotion recognition. Combining several audio features also improves the accuracy and resilience of the model. In conclusion, by giving AI-driven communication an emotional component, SER systems have the potential to completely transform human-machine interaction. These systems will develop into more responsive, inclusive, and human-centric systems as we keep refining SER using sophisticated models and a variety of data. They are just starting to play a part in the healthcare, education, and entertainment sectors, and further integration with multilingual data and real-time systems will open even more opportunities.

References

- [1]. Livingstone, S. R., & Russo, F. A. (2018). 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(5), e0196391.
- [2]. Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech*

Appendix C: System Deployment Considerations

- Hardware Requirements: At least 8GB RAM, GPU for training.
- Inference Time: ~100ms per sample on average.
- Scalability: Model can be quantified for deployment on edge devices.
- Security: Ensure privacy and encryption of voice data.

4.5. Upcoming Projects

- Application of self-supervised learning (e.g., HuBERT, Wav2Vec)
- Models that rely on attention
- Deployment in real time on embedded devices
- Multimodal facial and audio emotion recognition Show in Figure 3 & 4

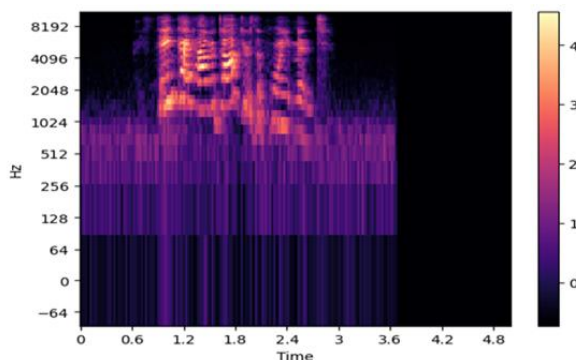


Figure 3 Angry

Communication, 48(9), 1162-1181.

- [3]. El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3), 572-587.
- [4]. Zhang, Z., Han, K., & Narayanan, S. (2019). Robust speech emotion recognition using DNN-based feature learning and ELM classifier. *ICASSP 2019 - IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645-6649.
- [5]. Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5200–5204.
- [6]. Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Networks*, 92, 60–68.
- [7]. Latif, S., Rana, R., Qadir, J., Epps, J., & Schuller, B. (2020). Deep architecture for multimodal speech emotion recognition: An overview. *IEEE Transactions on Affective Computing*, 12(3), 640-657.
- [8]. Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014). Learning salient features for speech emotion recognition using CNN. *IEEE Transactions on Multimedia*, 16(8), 2203-2213.
- [9]. Han, K., Yu, D., & Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. *Interspeech*, 223-227.
- [10]. Schuller, B., Steidl, S., & Batliner, A. (2009). The INTERSPEECH 2009 emotion challenge. *INTERSPEECH*, 312–315.