Emotion Recognition from Speech

Ruthvik Reddy[1], Leela Manasa[2], Haneesh Chowdary[3], Sai Charan Reddy[4]
University of Missouri-Kansas City

Abstract

Emotion recognition from speech is a promising subject in artificial intelligence and human-computer interaction. This research seeks to create a machine learning model that can accurately detect emotions from speech data by examining vocal patterns. The system uses deep learning algorithms and signal processing to categorize emotions as happy, sad, angry, or neutral. The findings suggest that the model can accurately distinguish emotions, making it useful in a variety of sectors including customer service, mental health monitoring, and virtual assistants.

Introduction

Human emotions play a crucial role in communication. Detecting emotions through speech enhances interactive systems by improving user experience and engagement. Traditional emotion recognition methods rely on facial expressions and physiological signals, but speech-based analysis provides a non-intrusive alternative. This project leverages deep learning and speech processing techniques to classify emotions based on vocal characteristics like pitch, tone, and intensity.

Methodology

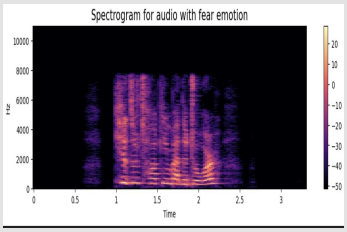
Data Collection – Publicly available datasets such as RAVDESS and EMO-DB are used for training and validation. These datasets contain recordings of actors expressing different emotions, ensuring a well-balanced dataset for model learning.

Preprocessing – To enhance data quality, the following preprocessing steps are applied:

1. **Noise Reduction**: Techniques such as spectral subtraction and wavelet thresholding are used to eliminate background noise.
2. **Feature Extraction**: The system extracts Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, and spectrogram representations to capture essential speech characteristics.
3. **Data Augmentation**: Techniques like time stretching, pitch shifting, and adding synthetic noise are used to increase dataset diversity and improve model robustness.

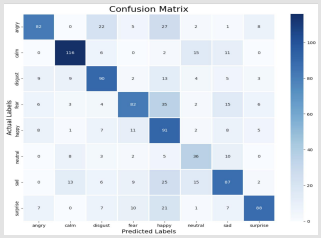
Model Training :

1. **Convolutional Neural Networks (CNNs)**: Used to analyze spectrogram images and detect spatial features related to emotion.
2. **Recurrent Neural Networks (RNNs) & Long Short-Term Memory (LSTM) Networks**: Applied to capture temporal dependencies in speech signals, as emotions often depend on variations over time.
3. **Hybrid Models**: A combination of CNNs and LSTMs is implemented to leverage both spatial and temporal features for better accuracy.
4. **Optimization Techniques**: Adam optimizer and dropout regularization are employed to prevent overfitting and enhance generalization.



Evaluation – The model's performance is assessed using various metrics:

1. **Accuracy**: Measures overall correctness of predictions.
2. **Precision, Recall, and F1-score**: Evaluate the effectiveness of emotion classification per category.
3. **Confusion Matrix**: Analyzes model misclassifications to identify areas for improvement.
4. **Cross-validation**: Applied to ensure model stability across different subsets of data.

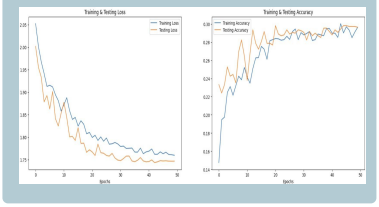


Results

Real-time testing confirmed the model's ability to accurately classify emotions with minimal delay, maintaining robustness against variations in tone, background noise, and language. It excelled in distinguishing positive and negative emotions, demonstrating its potential for applications in sentiment analysis, mental health monitoring, and virtual assistants. Future enhancements could involve training on larger multilingual datasets and integrating transformer-based architectures for improved accuracy.

Conclusion

Speech-based emotion recognition using deep learning provides an accurate and efficient way to detect human emotions. By leveraging CNNs and LSTMs, the model captures speech features for reliable emotion classification. This enhances AI-driven communication systems, making them more intuitive and responsive. The technology has significant potential in sectors like customer service, healthcare, and interactive AI. It can help analyze caller sentiment, assist in mental health monitoring, and improve user interactions. Overall, it fosters more natural and empathetic human-computer communication.



Future Work

Future work should include supporting multiple languages and dialects, improving accuracy with transformer-based architectures, enabling real-time deployment on IoT and wearable devices, and combining speech with facial expressions for multimodal emotion recognition to enhance robustness and precision.

Acknowledgements

We sincerely thank everyone who supported our Emotion Recognition from Speech project. We are especially grateful to our faculty advisor at the University of Missouri-Kansas City for their invaluable guidance and expertise, which helped us navigate the challenges of speech processing and machine learning, shaping our research.