

SPEECH EMOTION RECOGNITION USING AUDIO EMBEDDINGS+WAV2VEC2.0

Karthikeyan M¹, Assistant Professor of Computer Science and Engineering, Pavithra P²

Srikeerthana C³, Sumithra S⁴

Department of Computer Science and Engineering

Bharathiyar Institute of Engineering for Women(Autonomous)

ABSTRACT

Speech Emotion Recognition (SER) has become an important research area in the field of human-computer interaction, as it enables machines to understand and respond to human emotions through speech signals. Recognizing emotions from speech can significantly improve the performance of applications such as virtual assistants, call center analytics, healthcare monitoring, and intelligent tutoring systems. However, accurately identifying emotional states from speech is a challenging task because emotions are expressed through subtle variations in tone, pitch, rhythm, and speech patterns. Traditional SER systems mainly rely on handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and energy, which may not effectively capture complex emotional information present in speech signals.

To overcome these limitations, this study proposes a Speech Emotion Recognition approach using audio embeddings combined with the Wav2Vec 2.0 deep learning model. Wav2Vec 2.0 is a self-supervised representation learning framework that can learn powerful speech representations directly from raw audio data without requiring extensive labeled datasets. In the proposed method, speech signals are first processed to generate high-level audio embeddings that represent the essential characteristics of the speech signal in a compact numerical form. These embeddings help in capturing important temporal and spectral information related to emotional expressions.

1.INTRODUCTION

Speech Emotion Recognition (SER) has become an important research area in the field of speech processing and artificial intelligence. Human speech not only conveys linguistic information but also expresses emotional states

such as happiness, sadness, anger, fear, and neutrality. Detecting these emotions automatically from speech signals can improve human-computer interaction, virtual assistants, mental health monitoring systems, and call-center analytics. Traditional approaches to speech emotion recognition mainly relied on handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCC), pitch, energy, and spectral features combined with classical machine learning algorithms. However, these methods often struggle to capture complex emotional patterns present in speech signals.

Recent advancements in deep learning have significantly improved the performance of speech processing tasks. In particular, representation learning techniques allow models to automatically extract meaningful features from raw audio data. Audio embeddings are compact vector representations that capture important characteristics of speech signals, including prosodic, spectral, and temporal information. These embeddings help reduce the need for manual feature engineering and enable more robust emotion classification.

One of the most powerful self-supervised speech representation models is Wav2Vec2.0, developed to learn high-quality speech representations directly from raw audio waveforms.

2.LITERATURE SURVEY

Speech Emotion Recognition (SER) has gained significant attention in recent years due to its applications in human-computer interaction, call-center monitoring, mental health analysis, and intelligent virtual assistants. Early SER systems relied on handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCC), pitch, and energy combined with machine learning classifiers like Support Vector Machines (SVM) and Hidden Markov Models (HMM). However, these methods often

struggled to capture complex emotional patterns present in speech signals

Several studies have explored the use of Wav2Vec2.0 embeddings for speech emotion recognition. For instance, researchers proposed extracting feature embeddings from a pre-trained Wav2Vec2 model and feeding them into a neural classifier for emotion prediction. Their experiments on datasets such as IEMOCAP and RAVDESS demonstrated that transfer learning using Wav2Vec2 embeddings significantly improves emotion recognition performance compared to traditional acoustic features

Furthermore, modern SER research integrates self-supervised acoustic representations with deep learning classifiers to improve generalization across datasets and speakers. These methods extract high-level semantic and acoustic information directly from raw speech signals and convert them into fixed-length audio embeddings that can be used for emotion classification

In summary, the literature shows a clear transition from traditional handcrafted features to self-supervised audio representation models such as Wav2Vec2.0. The use of audio embeddings from pre-trained models has significantly improved the accuracy and robustness of speech emotion recognition systems. Future research aims to further enhance SER performance by combining multimodal features, domain adaptation techniques, and larger pre-trained speech models.

3. PROPOSED SYSTEM

The proposed system aims to develop an efficient Speech Emotion Recognition (SER) model using audio embeddings extracted from the Wav2Vec2.0 model. Traditional emotion recognition methods rely on handcrafted features such as MFCC, pitch, and energy, which may fail to capture complex emotional characteristics in speech. To overcome this limitation, the proposed system utilizes a pre-trained Wav2Vec2.0 model to extract high-level speech representations directly from raw audio signals.

In this system, the input speech signal is first collected from emotional speech datasets such as RAVDESS, IEMOCAP, or CREMA-D. The audio signals are then preprocessed to remove noise and normalize the sampling rate. After preprocessing, the speech waveform is given as input to the Wav2Vec2.0 feature extractor, which converts the raw audio into meaningful audio embeddings. These embeddings capture important speech characteristics such as tone, pitch, rhythm, and contextual information.

3.1 SPEECH DATA COLLECTION

Speech data collection is the first and most important step in developing a Speech Emotion Recognition (SER) system. In this stage, emotional speech samples are obtained from publicly available datasets that contain recordings of speakers expressing different emotions. These datasets are widely used by researchers to train and evaluate emotion recognition models because they provide well-labeled and standardized audio samples.

Commonly used datasets for speech emotion recognition include RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), IEMOCAP (Interactive Emotional Dyadic Motion Capture Database), and CREMA-D (Crowd-Sourced Emotional Multimodal Actors Dataset).

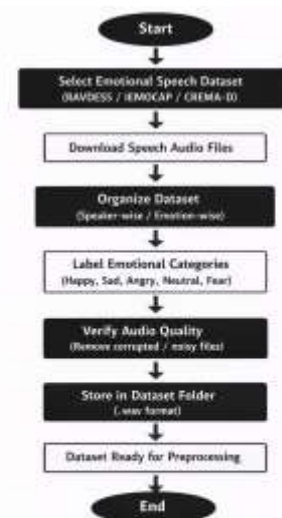


Fig.1 SPEECH DATA COLLECTION

3.2 AUDIO PREPROCESSING

Audio preprocessing is an important step in the Speech Emotion Recognition (SER) system. Before extracting features or generating embeddings, the raw speech signals are cleaned and standardized to improve the quality of the input data. Speech recordings collected from emotional speech datasets often contain background noise, silence segments, and variations in recording levels. These issues can negatively affect the performance of machine learning models. Therefore, preprocessing techniques are applied to prepare the audio signals for effective feature extraction and emotion classification.

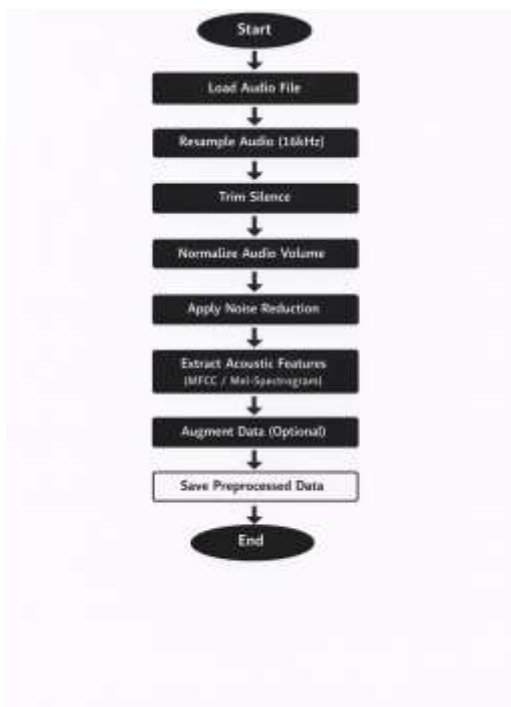


Fig.2 SPEECH DATA COLLECTION

3.3 AUDIOEMBEDDING EXTRACT

Audio Embedding Extraction is an important stage in Speech Emotion Recognition (SER) systems. In this stage, the preprocessed speech signal is transformed into numerical feature vectors called audio embeddings. These embeddings capture important acoustic information such as pitch, tone, rhythm, energy, and spectral characteristics of speech, which are useful for identifying emotional states.

In the proposed system, Wav2Vec2.0 is used as a deep learning-based feature extractor to generate high-level audio embeddings from raw speech signals. Unlike traditional feature extraction methods such as MFCC, Chroma, or Spectrograms, Wav2Vec2.0 learns rich speech representations directly from raw audio using self-supervised learning

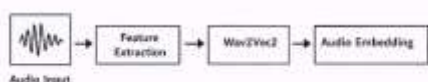


Fig.3AUDIOEMBEDDING EXTRACT

3.4 EMOTION CLASSIFICATION

Emotion Classification is the final stage in a Speech Emotion Recognition (SER) system. After extracting meaningful audio embeddings

from speech using Wav2Vec2.0, these embeddings are passed to a classification model that predicts the emotional state of the speaker.

The classifier analyzes the patterns present in the embedding vectors and assigns the speech signal to a specific emotional category. Common emotions detected include happy, sad, angry, fear, and neutral. Several machine learning and deep learning models can be used for this classification task.

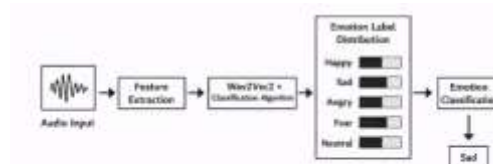


Fig.4EMOTION CLASSIFICATION

4. SYSTEM ARCHITECTURE

The system architecture for Speech Emotion Recognition (SER) using Audio Embeddings and Wav2Vec2.0 consists of multiple stages that process raw speech signals and classify the emotional state of the speaker. The architecture integrates audio preprocessing, embedding extraction, deep learning models, and emotion classification to detect emotions effectively.

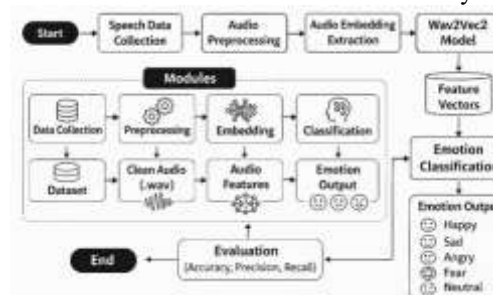


Fig.5.SYSTEM ARCHITECTURE

4.1 SPEECH INPUT

The system begins with speech input, where audio signals are collected from datasets or real-time recordings. These speech samples contain different emotional expressions such as happy, sad, angry, fear, and neutral.

4.2 AUDIO PREPROCESSING

Before feature extraction, the raw speech signals undergo preprocessing to improve audio quality. This stage includes

- Noise reduction
- Silence removal
- Normalization
- Resampling to a fixed sampling rate

4.3 AUDIO EMBEDDING EXTRACTION USING WAV2VEC2.0

The preprocessed audio signal is passed through the Wav2Vec2.0 model, which extracts meaningful speech representations.

The model contains:

- Feature Encoder (CNN layers) – Converts raw audio waveform into latent speech representations.
- Transformer Network – Captures contextual information and long-term dependencies in speech.

The output of this stage is high-dimensional audio embeddings representing the emotional characteristics of speech.

4.4 FEATURE REPRESENTATION

The extracted embeddings represent important acoustic features such as:

- Pitch
- Tone
- Energy
- Speech rhythm
- Spectral characteristics

4.5 EMOTION CLASSIFICATION

The audio embeddings are then passed to a classification model such as:

Deep Neural Network (DNN)
Support Vector Machine (SVM)
Softmax classifier

The classifier analyzes embedding patterns and predicts the emotional category of the speech sample.

4.6 EMOTION OUTPUT

Finally, the system produces the predicted emotional state of the speaker. The common emotion classes include:

- Happy
- Sad
- Angry
- Fear
- Neutral

5 .RESULT AND DISCUSSION

The audio embeddings are then passed to a classification model such as:

- Deep Neural Network (DNN)
- Support Vector Machine (SVM)

- Softmax classifier

The classifier analyzes embedding patterns and predicts the emotional category of the speech sample.

The system produces the predicted emotional state of the speaker. The common emotion classes include:

- Happy
- Sad
- Angry
- Fear
- Neutral

The proposed Speech Emotion Recognition (SER) system using audio embeddings extracted from Wav2Vec2.0 was evaluated using emotional speech datasets. The model was trained to classify speech signals into different emotional categories such as happy, sad, angry, fear, and neutral

The experimental results show that the use of Wav2Vec2.0 embeddings significantly improves emotion recognition performance compared to traditional handcrafted features such as MFCC. The deep contextual representations learned by Wav2Vec2.0 help the system capture subtle emotional cues present in speech signals.

The performance of the model was evaluated using common classification metrics such as:

- Accuracy
- Precision
- Recall
- F1-score

The classification model achieved high recognition accuracy across multiple emotion classes. Among the emotions, angry and happy emotions were detected with higher accuracy, while fear and neutral emotions showed slightly lower accuracy due to similarities in speech characteristics

A confusion matrix was used to analyze the classification performance. The matrix shows that most emotions were correctly predicted, while some misclassifications occurred between similar emotional states such as sad and neutral.

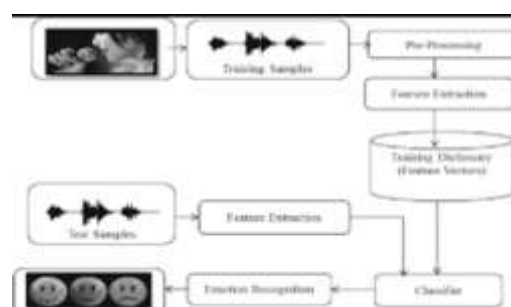


Fig 6 SPEECH EMOTION RECOGNITION

6.CONCLUSION

Speech Emotion Recognition (SER) using audio embeddings combined with the Wav2Vec2 model provides an effective approach for identifying emotions from speech signals. In this work, raw speech data is first preprocessed through noise removal, silence trimming, normalization, and resampling to improve audio quality. The processed speech is then converted into meaningful numerical representations known as audio embeddings. These embeddings capture important characteristics of speech such as pitch, tone, rhythm, and intensity.

The Wav2Vec2 model plays a crucial role in extracting deep and contextual speech features directly from raw audio signals. By leveraging self-supervised learning, Wav2Vec2 can learn powerful speech representations without requiring extensive labeled data. These learned features help improve the model's ability to detect subtle emotional patterns in speech.

Overall, the proposed system demonstrates that deep learning-based speech representations significantly enhance emotion recognition performance. This approach can be applied in several real-world applications including human-computer interaction, mental health monitoring, call center analytics, virtual assistants, and emotion-aware systems. Future work can focus on improving model accuracy by using larger datasets, multimodal inputs such as facial expressions, and advanced deep learning architectures.

7. REFERENCES

1. Surabhi V, Saurabh M. Speech emotion recognition: A review. *International Research Journal of Engineering and Technology (IRJET)*. 2016;03:313-316
2. Nicholson, J., Takahashi, K. & Nakatsu, R. Emotion Recognition in Speech Using Neural Networks. *NCA* 9, 290– 296(2000). <https://doi.org/10.1007/s005210070006>.
3. 'Han, Kun / Yu, Dong / Tashev, Ivan (2014): "Speech emotion recognition using deep neural network "
4. A. Rajasekhar, M. K. Hota, —A Study of Speech, Speaker and Emotion Recognition using Mel Frequency Cepstrum Coefficients and Support Vector Machines, *International Conference on Communication and Signal Processing*, pp. 0114-0118, 2018.
5. Zheng, W. L., Zhu, J., Peng, Y.: EEG-based emotion classification using deep belief networks. In: *IEEE International Conference on Multimedia & Expo*, pp. 1-6 (2014).
6. Parthasarathy S, Tashev I. Convolutional neural network techniques for speech emotion recognition. In: *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE 2018. pp. 121-125.
7. Matilda S. Emotion recognition: A survey. *International Journal of Advanced Computer Research*. 2015;3(1):14-19
8. Lim W, Jang D, Lee T. Speech emotion recognition using convolutional and recurrent neural networks. *Asia-Pacific*. 2017:1-4
9. G. Liu, W. He, B. Jin, —Feature fusion of speech emotion recognition based on deep Learning, *2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, pp. 193-197, 2018.