Robust Speech Micro-Emotion Detection using MFCC Features and CNN in Noisy Environments

Ujwalla Gawande[1], Hemalatha[2,]

[1]Global Postdoctoral & Researcher Programme, Lincoln University College Malaysia

[1]Professor, Department of IT, Yeshwantrao Chavan College of Engineering,

Nagpur, Maharashtra, India

ujwallgawande@yahoo.co.in

[2]Professor

Department of computer science and business systems

Panimalar Engineering college

Poonamallee. Chennai. Tamilnadu. India

pithemalatha@gmail.com

Abstract: Micro-emotions expressed through speech are subtle emotional cues that can be identified from a person's vocal communication. Detecting these emotions is a challenging task because they occur within very short time intervals and often exhibit minimal acoustic variations. This study presents an approach for automatic micro-emotion recognition from speech signals. The proposed framework begins with the application of a Voice Activity Detection (VAD) algorithm to identify and eliminate non-speech portions from the audio recordings. The resulting speech signal is then segmented into smaller units for further analysis. Acoustic features, including Root Mean Square (RMS) Energy, Pitch, Mel-Frequency Cepstral Coefficient (MFCC) Mean, MFCC Variance, and Zero Crossing Rate (ZCR), are extracted from each segment to capture the underlying characteristics of speech.

To improve the effectiveness of emotion classification, the most informative features are selected using the ReliefF feature selection algorithm. These optimized features are subsequently utilized for micro-emotion recognition. The extracted emotional patterns are mapped to corresponding micro-level emotional categories and used to train a Convolutional Neural Network (CNN)-based classification model. The proposed methodology has been evaluated using the CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset), which contains 7,442 original video clips. Experimental results demonstrate the capability of the system to accurately identify micro-level emotions, even in the presence of background noise, indicating its potential for robust real-world speech emotion recognition applications.

Key words: VAD, Speech micro emotion, MFCCs.

1. Introduction

In today's era, understanding someone's true emotion is very important. Emotions can be expressed in different forms like text, speech, facial expressions, or body language. Speech is a basic mode of communication for humans. Humans share most of its emotions through Speech. Mostly they are natural and its expression is closely associated with feelings. Generally, gender and culture differences have limited influence on the appearance of micro-expressions, however individual's background, context, and situational factors can affect more on how these expressions are exhibited.

The human emotions are broadly classified into two classes: Comfortable Emotions and Uncomfortable Emotions, i.e. positive and negative emotion, respectively as shown in figure 1. Each of these categories are further subdivided into macro/core level emotions such as Happy, Loved, Confident, Playful, Embarrassed, Angry, Scared and Sad. These core emotions form the foundational states. Those macro level emotions are further categorized into micro emotions which are more specific emotional expressions derived from the major states.

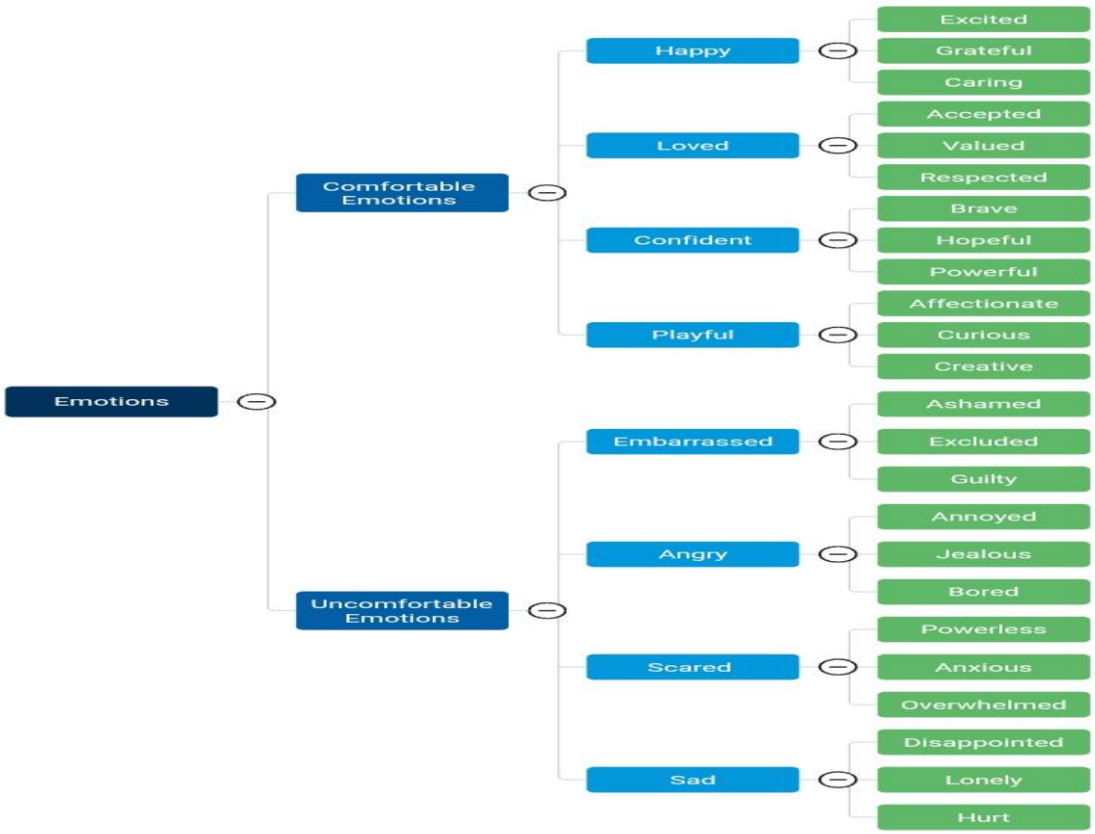

Figure 1. Categorization of emotions at macro and micro level

The work observed in literature mostly focuses on macro level emotion detection. However very less work has been done on micro level. Developing micro-expressions-based emotion recognition system, can significantly help us understand one's emotions and psychological behaviour better and can be useful in the future for understanding psychology of a person. Hence, speech micro-emotion detection is an important area of research in artificial intelligence. Such systems have the potential to support advanced human–computer interaction by enabling intelligent machines and robots to accurately perceive, interpret, and respond to human emotions in a more natural and effective manner. Hence this paper proposed a micro-expressions detection using relief and Root Mean Square (RMS), Mel-frequency Cepstral Coefficients (MFCCs), Zero crossing Rate (ZCR) and pitch. Those features are mapped to the micro expression and are trained using CNN. The results of the proposed approach outperform the existing approaches in literature.

2. Related Work

In today's era, humans' mental health is very important. For the same detecting an emotion at right time may be very useful for avoiding any mishaps in future. In the literature, much research has been carried out by the researchers for detecting the macro emotion from face. Very few works have been carried out on micro level emotion detection. For working on that emotion, the various datasets available are MSP-IMPROV [6], Emo-DB [8], TESS [10], RAVDESS [9], etc. Speech is having a rich source of emotion. Speech Emotion Recognition (SER) investigates features such as tone, rhythm, and pitch which carry critical information. Traditional SER methods employed handcrafted features such as MFCCs and ZCR to represent speech signals and facilitate emotion detection. Deep learning replaced these hand-crafted techniques with data-driven representations that capture subtle emotional messages.

Recently work introduce by Yang et al. [1] developed a dual-layer LSTM model for improvement of the emotion detection from raw speech. They claimed the improved accuracy of 2% over standard LSTM baselines on the RAVDESS dataset. Yuanyuan [2] converted raw speech data in the IEMOCAP dataset into spectrogram images and extracted feature maps using AlexNet. The feature maps processed to highlight the essential feature channels in the attention layer. In this model they used Grad CAM.

Most of the work in this area is focussed on the models of Deep learning. Techniques such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and attention mechanisms have significantly enhanced the accuracy and efficiency of SER systems. CNNs have been used to extract high-level features from spectrograms, which are then used for emotion classification [3]. RNNs are used to process sequential data, making them suitable for time-series analysis, such as speech signals [4]

From this literature, it has been observed that most of the work has been performed on macro level emotion detection and very few on micro level. Hence there is a need of micro level emotion detection using speech, to carefully understand the persons emotions such as lonely, annoyed, guilty, etc. This will help in taking necessary/corrective action in time and improve the life style.

3. Methodology

The proposed approach of micro-expression detection system is shown in figure 2. It consists of the steps like preprocessing, feature extraction, micro emotion mapping and training the model based on those mappings. Here the audios from the CREMA-D data set have been given as an input to the proposed system.

3.1 pre-processing

Normally when we have audio clips for micro emotion detections, it contains the background noises. So, we need to remove these background noises from audio clips. Voice Activity Detection (VAD) algorithm has been used to separate audio into speech and non-speech sections. This technique works by transforming an audio signal using Fast Fourier Transform (FFT), calculating a noise threshold for each frequency band, and suppress noise components. Now we are left with only speech information in an audio. Those speech signals are segmented into smaller frames of 25 ms with the help of a python

library librosa. These segments detect small movement of speech which helps in the further step of feature extraction and allow analysis of speech features at micro level.

3.2 Feature Extraction

Acoustic features that carry emotional variations are extracted from each frame for micro-emotion classification. These features capture changes in voice frequency, loudness, spectral content, voice quality, and vocal tract characteristics, thereby providing a numerical representation of the underlying speech characteristic. For each frame the following acoustic features are computed.

- Root Mean Square (RMS) Energy: Measures the intensity or loudness of speech and captures energy fluctuations.
- Mel-Frequency Cepstral Coefficients (MFCCs): Characterize the spectral properties of speech by representing the vocal tract response and perceptual aspects of sound. Statistical measures such as MFCC mean and variance are computed to capture temporal variations.
- Zero-Crossing Rate (ZCR): Quantifies the rate of signal polarity changes and reflects signal fluctuations and speech dynamics.
- Pitch: The variation in voice frequency is captured

3.2.1. Energy (RMS): For each segmented part of the frame in speech signal, the RMS energy is calculated using equation 1. This energy captures the short intensity variations in speech signal that further serves as an important acoustic feature for micro-emotion recognition.

$$E = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x[n]^2} \qquad \text{Equation 1}$$

Where: E stands for Energy

$x[n]$= Intensity of audio signal sample

N = Number of samples in each frame

This results in energy value for the segmented part, hence each frame, resulted in sequence of energy values in speech at different times in audio.

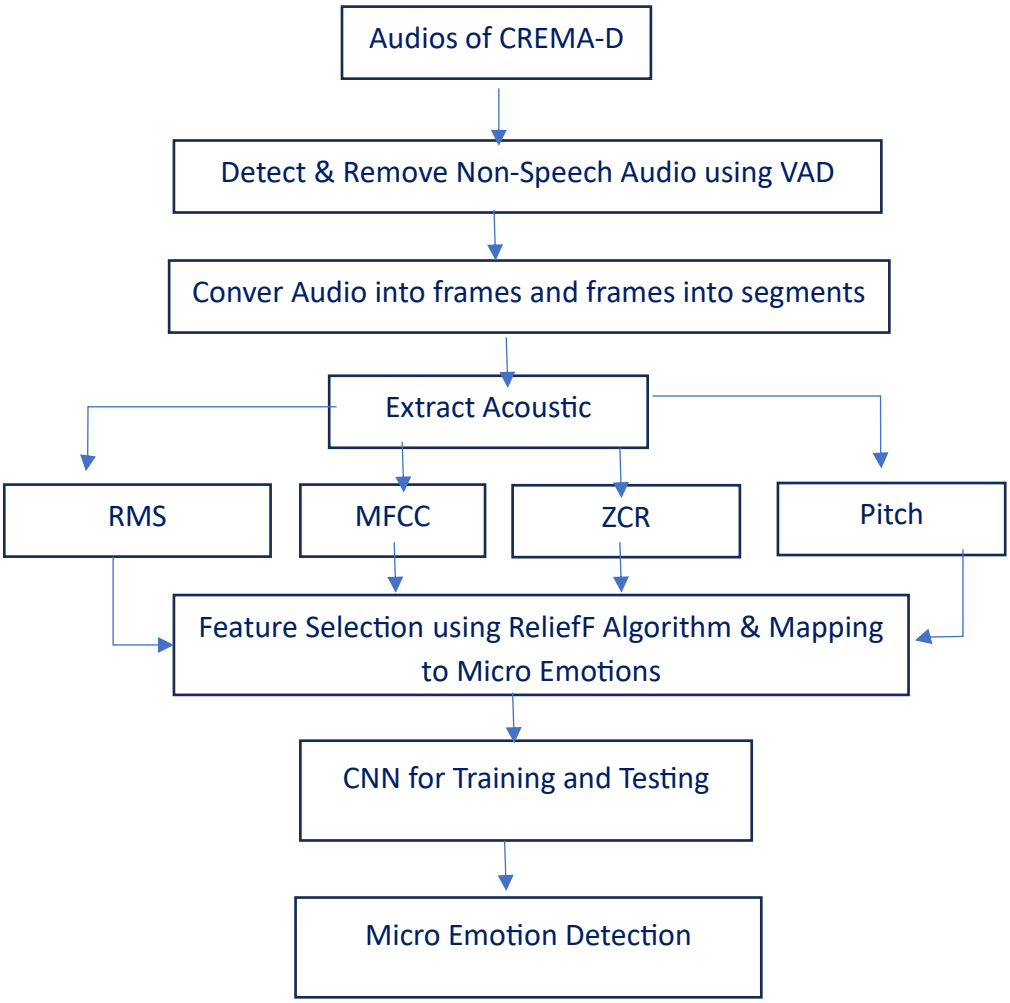

Figure 2: Block diagram for micro level emotion detection

3.2.2 MFCC Mean and variance: Humans do not perceive all sound frequencies equally, instead they are more sensitive to lower frequency. For each frame, the FFT is calculated and the bank of Mel-scale filters are used to compress the high frequency. For the remaining spectrum the logarithmic is found and then the DCT is applied to get the MFCC coefficients. These coefficients represent the important resonance features of voice for recognition. For each frame Y suppose X MFCC coefficients are extracted, it gives Y x X features. Different speech samples may have different features size. For further training the classifier, we require fixed-length feature vector. Hence, we calculate MFCC Mean and variance, MFCC Mean will average values across frames and MFCC variance will give variations value across the frames.

$\mu_{MFCC} = \frac{1}{T}\sum_{t=1}^{T} C_t$ Equation 2

$$\sigma^2_{MFCC} = \frac{1}{T}\sum_{t=1}^{T}(C_t - \mu_{MFCC})^2$$ Equation 3

Where: $\mu_{MFCC} - MFCC\ Mean$

$\sigma^2_{MFCC} - MFCC\ variance$

C_t = MFCC coefficient at frame t

T = Total frames

3.2.3. ZCR (Zero Crossing Rate): Speech signal has different signs like positive and negative. The rate at which these signals change their sign is found by ZCR i.e., from positive to negative or vice versa. It represents the content in the frequency and noise of the signal.

$$ZCR = \frac{1}{N-1}\sum_{n=1}^{N-1}\mathbb{1}(x[n]\cdot x[n-1] < 0) \quad \text{Equation 4}$$

Where: $\mathbb{1}$= Indicator function (1 if condition true, else 0)

(x(n)): Speech sample amplitude

3.2.4. Pitch: It finds the vibration rates in the speech, measured in Hz. It measures the similarity between a voice signal and its delayed version. For a speech frame x[n] of length N

$$AC(k) = \frac{1}{N-1}\sum_{n=0}^{N-1-\mathrm{k}}(x[n]x[n+\mathrm{k}]) \quad \text{Equation 5}$$

where:

$AC(k)$ = autocorrelation at lag k

x[n] = speech sample at time n

N = number of samples in the frame

k = lag

The pitch period P0 is found using lag corresponding to the first significant peak

$$P0 = \frac{kmax}{fs} \quad \text{Equation 6}$$

where:

- kmaz= lag at the maximum autocorrelation peak
- fs = sampling frequency (Hz)

Finally, the fundamental frequency is

$$FF = \frac{1}{P0} \quad \text{Equation 7}$$

These acoustic descriptors collectively encode emotional cues present in speech and serve as input features for subsequent feature selection and deep learning-based micro-emotion classification. With the help of above Equations (1-7) we calculate the acoustic feature. Now we use ReliefF algorithm to screen these high-dimensional features, selecting the most relevant ones to reduce redundancy and computational complexity. By selecting features with higher weights, it helps to creates a refined feature subset that improves the efficiency of the subsequent classifier. Further mapping of those features is done to core emotions and then micro emotions as shown in Table 1.

Table 1: Micro and core emotion mapping

Core Emotion	**Micro Emotion**	**Energy (RMS)**	**Pitch**	**MFCC**	**ZCR**	**Speech-Based Features**
Sad	Hurt	0.15–0.25	100–140	0.05–0.10	0.02–0.05	Low energy, smooth MFCC, low variance

Scared	Overwhelmed	0.60–0.75	200–250	0.70–0.90	0.50–0.70	Highly irregular MFCC, high variance
Angry	Annoyed	0.50–0.65	160–200	0.30–0.50	0.25–0.40	Moderate MFCC variation, medium energy
Embarrassed	Guilty	0.20–0.30	110–140	0.15–0.25	0.05–0.10	Low energy + irregular MFCC variation
Playful	Creative	0.55–0.70	190–220	0.40–0.60	0.30–0.40	Dynamic MFCC patterns, moderate-high variance
Confident	Brave	0.70–0.85	190–220	0.30–0.40	0.20–0.30	Stable MFCC + strong consistent energy

Now we have results of mapping. We use these mapping to train Convolutional Neural Network (CNN). CNN is an algorithm used to recognize patterns in data. With repeated training we get a proper class for micro emotion. During testing the CNN is able to classify the micro emotions. With this our approach achieved a higher solidify accuracy.

4. Result and discussion

As an outcome of the proposed approach, we have received an accuracy of 85% at micro level emotion detection. Further a comparison with the existing approaches have also been shown in Table 2. The same has been depicted in figure 3. From this it has been observed that, very less work has been performed on micro level emotion detection. In some cases, the mixed emotions are also detected, however they have not covered all the micro level emotions.

Table 2. Comparison analysis of proposed methodology with existing approaches

Macro/Micro	**Algorithm / Model**	**Dataset**	**Size of the dataset**	**Performance / Accuracy**	**Reference**
Micro	Proposed algorithm	Crema-D	7,442 videos	85%	Proposed methodology
7 Emotions Macro	CNN	Emo-DB	535 audios	90.36%	Kaur & Kumar [8]

From the above results, we can claim that our proposed algorithm outperforms the existing approaches particularly micro emotions classification in speech with the Accuracy of 85%.

5. Conclusion

This paper presents an approach for micro level emotions detection in audios. Here firstly we use VAD algorithm to detect & remove non-speech from audio. Those signals are converted into frames and further small segments, with the help of librosa python library. Three techniques are used for calculating the acoustic features viz, RMS, MFCC, pitch and ZCR. Further ReliefF Algorithm has been used to choose the most relevant feature to detect micro emotion, which is further mapped to the micro level emotions. Based on this mapping the CNN model is trained. The work has been performed on CREMA-D data set consisting of 7,442 original videos. The results have shown the accuracy of 85%. This approach is very useful for real time scenarios to understand the emotions of a person. The future work lies in increasing the accuracy of the proposed framework by using multiple modalities.

6. References:

[1] X. Yang, S. Yu, and W. Xu, “Improvement and Implementation of a Speech Emotion Recognition Model Based on Dual-Layer LSTM,” arXiv preprint arXiv:2411.09189, Nov. 2024, doi: 10.48550/arXiv.2411.09189.

[2] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, “Attention Based Fully Convolutional Network for Speech Emotion Recognition,” arXiv preprint arXiv:1806.01506v2 , May 2019.

[3] A. Mukhamediya, S. Fazli, and A. Zollanvari, “On the Effect of Log-Mel Spectrogram Parameter Tuning for Deep Learning-Based Speech Emotion Recognition,” IEEE Access, vol. 11, pp. 61950–61957, Jun. 2023, doi: 10.1109/ACCESS.2023.3287093.

[4] B. Akçay and K. Oguz, “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,” Speech Communication, vol. 116, Jan. 2020, doi: 10.1016/j.specom.2019.12.001.

[5] G. Alhussein, I. Ziogas, S. Saleem, and L. J. Hadjileontiadis, “Speech emotion recognition in conversations using artificial intelligence: a systematic review and meta-analysis,” Artificial Intelligence Review, vol. 58, no. 7, Apr. 2025, doi: 10.1007/s10462-025-11197-8.

[6] L. Pepino, P. Riera, and L. Ferrer, “Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings,” arXiv preprint arXiv:2104.03502v1, Apr. 2021.

[7] S. Somvanshi, A. Meher, P. Hagawane, M. Adhav, and J. Harne, “Multilingual Speech Emotion Recognition Using Deep Learning,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 10, no. 8, pp. 183–189, Nov.–Dec. 2024, doi: 10.32628/CSEIT2410772.

[8] J. Kaur and A. Kumar, “Speech Emotion Recognition Using CNN, k-NN, MLP and Random Forest,” in Computer Networks and Inventive Communication Technologies, pp. 499–509, Jun. 2021, doi: 10.1007/978-981-15-9647-6_39.

[9] S. U. Bhandari, H. S. Kumbhar, V. K. Harpale, and T. D. Dhamale, “On the Evaluation and Implementation of LSTM Model for Speech Emotion Recognition Using MFCC,” in Computer Networks and Inventive Communication Technologies, Springer International Publishing, pp. 421–434, Jan. 2022, doi: 10.1007/978-981-16-7182-1_33.

[10] V. M. Praseetha and P. P. Joby, “Speech emotion recognition using data augmentation,” International Journal of Speech Technology, vol. 25, pp. 783–792, Aug. 2021, doi: 10.1007/s10772-021-09883-3.