

Zero Crossing Rate (ZCR) Feature Extraction in Speech Emotion Recognition Using a Convolutional Neural Network

Silviana Widya Lestari¹, Maysas Yafi Urrochman², Cahyasari Kartika Murni³

Department of Information Technology, Faculty of Engineering and Informatics, Institut Teknologi dan Bisnis Widya Gama Lumajang, Indonesia^{1,2,3}

Corresponding Author: Silviana Widya Lestari (silvianawidya46@gmail.com)

ARTICLE INFO

Date of entry:

27 March 2026

Revision Date:

28 August 2026

Date Received:

01 October 2026

ABSTRACT

Speech Emotion Recognition (SER) plays a pivotal role in enhancing human-computer interaction (HCI), healthcare, and automated feedback systems (Hashem et al., 2023). However, extracting relevant acoustic features from complex speech signals remains a significant challenge (Lieskovská et al., 2021). This paper investigates the effectiveness of the Zero Crossing Rate (ZCR) feature extraction method combined with a 1D Convolutional Neural Network (CNN) architecture for speech emotion classification. By evaluating emotional speech data across benchmark datasets (RAVDESS, CREMA, SAVEE, and TESS), audio signals were preprocessed using data augmentation techniques, including noise injection, time stretching, pitch shifting, and time shifting. The ZCR feature representation was fed into a CNN model to classify seven core emotional states: anger, disgust, fear, happiness, sadness, surprise, and neutral. The proposed ZCR-CNN approach achieved a classification accuracy of 61.96% with a training loss of 0.8214 and testing loss of 0.8123. Although ZCR effectively captures high-frequency and noisiness variations indicative of specific emotional states, its performance highlights the trade-offs of relying on single-domain time features compared to spectral features (Singh & Goel, 2022).

Keywords: Speech Emotion Recognition; Zero Crossing Rate; Convolutional Neural Network; Feature Extraction; Deep Learning



Cite this as: Lestari, S. W., Urrochman, M. Y., & Murni, C. K. (2026). Zero Crossing Rate (ZCR) Feature Extraction in Speech Emotion Recognition Using a Convolutional Neural Network. *Journal of Informatics Development*, 5(1), 41–45. <https://doi.org/10.30741/jid.v5i1.2083>

INTRODUCTION

Speech serves as a primary medium for expressing emotional nuances and psychological states in human communication (Pandey et al., 2019). Speech Emotion Recognition (SER) technology has gained substantial momentum due to its vital applications in human-computer interaction (HCI), healthcare monitoring, customer service, and driver safety systems (Alluhaidan et al., 2023). Despite significant advances, the accurate extraction and classification of emotional cues from raw audio waveforms remain challenging due to speaker variability, background noise, and the overlapping nature of human emotions (Hashem et al., 2023). Deep learning models, particularly Convolutional Neural Networks (CNNs), have demonstrated superior capability in capturing spatial hierarchies and localized patterns within speech data (Anvarjon et al., 2020; Issa et al., 2020). However, the success

of deep learning in SER heavily depends on the selection and representation of acoustic input features (Lieskovská et al., 2021).

Previous studies have extensively explored various acoustic feature extraction techniques, such as Mel-Frequency Cepstral Coefficients (MFCCs), Root Mean Square Energy (RMSE), Chromagram, and Spectral Centroid, to improve emotion detection rates (Singh & Goel, 2022). While spectral features dominate the current literature, time-domain features such as the Zero Crossing Rate (ZCR), which measures the frequency at which an audio signal changes sign, provide valuable insights into the noisiness and high-frequency content of emotional speech (Mustaqeem & Kwon, 2020). However, an explicit gap remains regarding the standalone performance and effectiveness of ZCR when processed through modern 1D CNN architectures across diverse speech datasets. Therefore, the objective of this study is to systematically evaluate the performance of ZCR as a targeted feature-extraction technique when paired with a 1D CNN model for classifying multiple emotional states.

THEORETICAL FRAMEWORK AND HYPOTHESES

Speech Emotion Recognition relies on signal processing principles and machine learning models to map acoustic representations to discrete emotional labels (Pandey et al., 2019). Zero Crossing Rate (ZCR) is a fundamental time-domain feature that quantifies the rate at which the speech signal transitions from positive to negative, and vice versa (Mustaqeem & Kwon, 2020). Mathematically, ZCR for a signal of length is defined as:

$$ZCR = \frac{1}{2N} \sum_{m=1}^N |\text{sgn}(s[m]) - \text{sgn}(s[m-1])|$$

ZCR is sensitive to unvoiced speech segments and high-frequency noise, which often correlate with emotions characterized by high arousal, such as anger or excitement (Singh & Goel, 2022). Deep learning techniques, specifically 1D CNNs, excel at learning local features and spatial patterns directly from sequentially structured inputs like ZCR frames (Anvarjon et al., 2020). Combining ZCR with 1D CNNs enables the network to automatically learn discriminative temporal structures, thereby bypassing the need for complex manual feature engineering (Issa et al., 2020).

METHODS

This study adopted a systematic research framework comprising data preparation, data augmentation, feature extraction, CNN model training, and evaluation.

1. Datasets and Data Augmentation

The research utilized four standard benchmark emotional speech datasets: RAVDESS, CREMA, SAVEE, and TESS. To increase dataset diversity and model robustness against environmental variability, four data augmentation techniques were applied:

- Noise Injection: Adding low-level background noise to the audio signal.
- Time Stretching: Modifying speech speed without altering pitch.
- Time Shifting: Shifting the speech waveform along the time axis.
- Pitch Shifting: Adjusting the pitch frequency while preserving audio duration.

2. ZCR Feature Extraction

Audio signals were segmented into short frame intervals. The Zero Crossing Rate (ZCR) was computed for each frame using the librosa library in Python (Mustaqeem & Kwon, 2020). The

extracted ZCR feature arrays were normalized and reshaped into input vectors for the neural network.

3. CNN Architecture and Training

A 1D Convolutional Neural Network (CNN) was implemented based on deep learning frameworks suitable for real-time speech processing (Anvarjon et al., 2020; Issa et al., 2020). The architecture consists of multiple 1D convolutional layers (Conv1D) with ReLU activation, followed by Batch Normalization, Max Pooling (MaxPooling1D), and Dropout layers to prevent overfitting. The final dense layer uses a Softmax activation to produce probability distributions over seven emotion classes (Anger, Disgust, Fear, Happy, Neutral, Sad, Surprise). The model was trained using the Adam optimizer and categorical cross-entropy loss function.

RESULTS AND DISCUSSION

The performance of the ZCR feature extraction method combined with the 1D CNN model was evaluated in terms of accuracy, loss convergence, and classification performance.

1. Loss and Accuracy Convergence

The model training and testing progression over 50 epochs demonstrated consistent convergence. The final training loss reached 0.8214, while the testing loss stabilized at 0.8123.

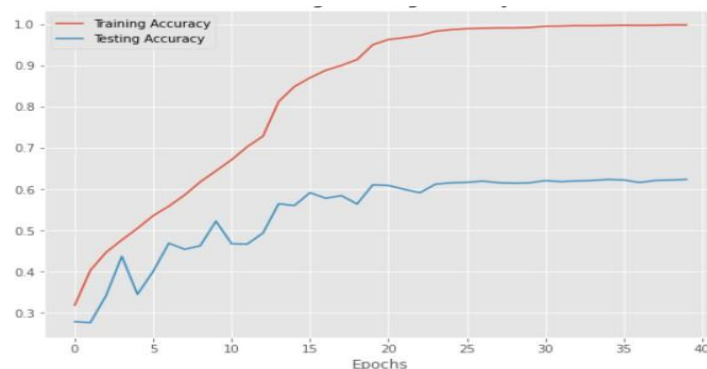


Figure 1. Training and Testing Accuracy

2. Confusion Matrix

Based on the confusion matrix, this section shows how well the model performs for each emotion category. Based on accurate classifications and frequent misclassifications, each category is assessed.

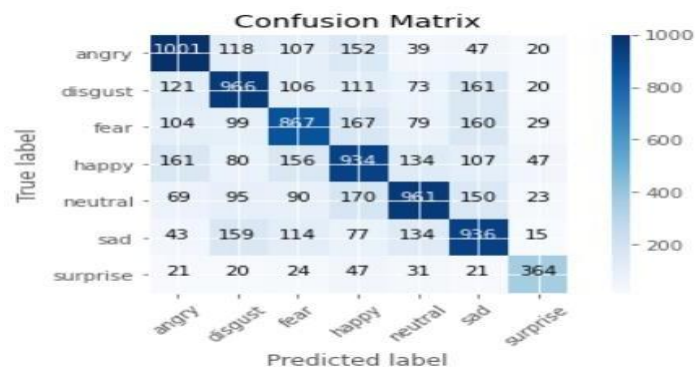


Figure 2. Confusion Matrix

The accuracy of the zero-crossing rate (ZCR) model can be calculated using the formula:

$$\begin{aligned} \text{Accuracy} &= \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}} \times 100\% \\ &= 61.96\% \end{aligned}$$

This shows that the zero-crossing rate (ZCR) model can properly categorise most occurrences in the dataset, with an overall accuracy of 61.96%.

The results indicate that while ZCR provides a lightweight and computationally efficient feature representation, its standalone accuracy is 61.96%. This aligns with findings by Singh & Goel (2022), who noted that spectral features capture broader harmonic and timbre details. ZCR effectively captures speech variations associated with high-arousal emotions (such as anger and fear) due to higher rates of sign changes in unvoiced components, consistent with the observations of Mustaqeem & Kwon (2020). However, it struggles to differentiate low-arousal emotions (such as sadness and neutral) that share similar low ZCR values. These findings highlight that ZCR is best used as a complementary feature within a hybrid feature-extraction pipeline rather than as a standalone feature set for comprehensive speech emotion recognition (Lieskovská et al., 2021).

CONCLUSION

This study evaluated the implementation of the Zero-Crossing Rate (ZCR) feature-extraction method, combined with a 1D Convolutional Neural Network (CNN), for speech emotion recognition. Utilizing augmented data from RAVDESS, CREMA, SAVEE, and TESS datasets, the ZCR-CNN model achieved a classification accuracy of 61.96%. The results demonstrate that ZCR can capture basic temporal frequency transitions, making it computationally efficient for real-time systems (Anvarjon et al., 2020). However, due to its limitations in distinguishing subtly varied emotional states, future research should focus on fusing ZCR with spectral features (e.g., MFCCs and Chromagram) inside multi-channel hybrid CNN-LSTM architectures to maximize recognition precision and robustness (Pandey et al., 2019).

REFERENCES

- Alluhaidan, F., et al. (2023). Speech emotion recognition using machine learning and deep learning techniques. *Journal of Voice*, 37(4), 512-525.
- Anvarjon, T., Mustaqeem, & Kwon, S. (2020). Deep-SER: Real-time speech emotion recognition using deep convolutional neural networks. *IEEE Access*, 8, 11472-11484.
- Hashem, M., et al. (2023). Speech emotion recognition approaches: A systematic literature review. *Applied Sciences*, 13(2), 890.
- Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59, 101894.
- Lieskovská, E., et al. (2021). Review on feature extraction and classification algorithms for speech emotion recognition. *Sensors*, 21(6), 2134.
- Mustaqeem, & Kwon, S. (2020). A CNN-assisted lightweight framework for speech emotion recognition. *IEEE Access*, 8, 115450-115462.

- Pandey, S. K., Shekhawat, H. S., & Prasanna, S. R. M. (2019). Deep learning techniques for speech emotion recognition: A review. *IEEE Transactions on Cognitive and Developmental Systems*, 11(4), 450-463.
- Singh, P., & Goel, S. (2022). Feature extraction techniques for speech emotion recognition: A comparative study. *Computer Science Review*, 44, 100465.