

Indonesian speech emotion recognition: feature extraction and neural network approaches

Izza Nur Afifah¹, Tri Budi Santoso², Titon Dutono³

¹Department of Informatics and Computer Engineering, Politeknik Elektronika Negeri Surabaya, Surabaya, Indonesia

²Department of Creative Multimedia Technology, Politeknik Elektronika Negeri Surabaya, Surabaya, Indonesia

³Department of Electrical Engineering, Politeknik Elektronika Negeri Surabaya, Surabaya, Indonesia

Article Info

Article history:

Received Aug 31, 2024

Revised Mar 26, 2025

Accepted May 24, 2025

Keywords:

Cohen's Kappa

Convolutional neural networks

Long short-term memory

Mel-frequency cepstral coefficients

Speech emotion recognition

ABSTRACT

This study explored the challenges of emotion recognition in Indonesian speech using deep learning techniques, addressing the complex nuances of emotional expression in spoken language that posed significant difficulties for automatic recognition systems. The research focused on the application of feature extraction methods and the implementation of convolutional neural networks (CNN) and a hybrid convolutional neural networks-long short-term memory (CNN-LSTM) model to identify emotional states from speech data. By analyzing key features of speech signals, including mel frequency cepstral coefficient (MFCC), zero crossing rate (ZCR), root mean square energy (RMSE), pitch, and spectral centroid, the study evaluated the models' ability to capture both spatial and temporal patterns in the data. Testing was conducted using an Indonesian dataset comprising 200 samples. The CNN model, utilizing four features (MFCC, ZCR, RMSE, and pitch), and the CNN-LSTM model, which used three features (MFCC, ZCR, and RMSE), both achieved an emotion classification accuracy of approximately 88%. The result showed that the CNN-LSTM model achieved comparable performance with a simpler feature set compared to the CNN model. This highlighted the significance of choosing the appropriate techniques in feature extraction and classification to enhance the accuracy of identifying emotions from speech data while also managing computational complexity.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Tri Budi Santoso

Department of Creative Multimedia Technology, Politeknik Elektronika Negeri Surabaya

Jalan Raya ITS, Keputih, Sukolilo, Surabaya, East Java 60111, Indonesia

Email: tribudi@pens.ac.id

1. INTRODUCTION

Speech communication serves as the simplest and effective approach that people have in order to communicate information. The importance of speech becomes evident when alternative communication methods, such as text messages or emails, are commonly used but can easily be misinterpreted. When we attempt to express emotions in writing, emojis often become necessary aids in text messaging [1]. Thus, speech is the most effective method to communicate in human life, as it carries a wealth of information through both linguistic and paralinguistic elements [2].

The advancement of information and communication technology (ICT) technology has opened up new possibilities for how humans interact with computers. Given that understanding emotional states enhances interpersonal comprehension, there is a need to integrate this concept into computer systems. This idea inspired the establishment of speech emotion recognition (SER), a field focused on identifying and interpreting emotional states conveyed through speech. Many studies have been conducted to explore SER,

but the topic still presents significant challenges. SER technology has potential uses across several fields, including healthcare, call centers, and education [3]–[5]. In healthcare, it can help in the diagnosis of psychological problems like depression, autism, and other mental disorders. In call centers, it helps measure customer satisfaction. In education, particularly in distance learning, it can enhance the learning experience. Despite its significant potential, challenges remain, such as the lack of diverse datasets, choosing the right features, and the choice of effective intelligent recognition techniques [6]–[8].

The majority of SER research has focused on languages with abundant resources and widespread use, such as English or German [9]–[11]. Although these studies have deepened our understanding of detecting emotions in speech, there remains a considerable gap in exploring resource-limited languages like Indonesian. In recent years, research on emotion detection in Indonesian speech has begun to emerge, covering areas such as emotion detection in films [12], recognition using acoustic and lexical features [13], and automatic emotion recognition [14]. Despite Indonesian being spoken by over 200 million people, research attention in SER remains limited. The scarcity of corpora and standardized databases hampers the progress of SER research in Indonesian. Cross-lingual emotion recognition experiments have been conducted due to these limitations [15].

In simple terms, SER consists of two primary components: feature extraction and classification [16], [17]. Feature extraction involves identifying characteristics related to emotion within speech signals [18]. The goal is to extract emotional information from spoken language by converting the raw speech signals into relevant feature sets. SER frameworks divide characteristics into four categories: prosodic features, spectral features, voice quality features, and Teager energy operator (TEO)-based features [2]. The challenge lies in choosing the most essential features that are able to differentiate between different emotions [19]. Mel-frequency cepstral coefficients (MFCC) is effective in capturing important spectral characteristics based on human perception of frequency, making it relevant for detecting spectrum changes associated with emotions. zero crossing rate measures how frequently the value of the audio signal changes from above to below zero, providing information about the temporal aspects that may change with emotion. Root mean square energy (RMSE) measures the average energy of the speech signal, which can reflect varying sound intensity levels associated with emotions. Pitch measures the fundamental frequency of the speech, where changes in pitch are often linked to emotional variation. Spectral Centroid measures the average frequency location within the spectrum, reflecting the brightness of the sound, which may change with different energy distributions due to emotions [20]–[24].

Classification is the second crucial step in SER. It involves applying machine learning models to the extracted features to identify the emotions expressed in speech. There are two main approaches to SER classification: conventional classifiers and deep learning classifiers. Recent developments indicate that problems in SER are being addressed with more emphasis on machine learning techniques, especially deep learning approaches. Deep learning methods have demonstrated significant improvements in emotion recognition, offering advantages such as scalability, parameter tuning, and customizable functions [2]. Several researchers have explored various neural network methodologies, including artificial neural networks (ANN), convolutional neural networks (CNN), deep neural networks (DNN), recurrent neural networks (RNN), and long short-term memory (LSTM) [25]–[28]. CNN and LSTM are increasingly recognized for SER tasks because they effectively capture temporal dependencies and spatial patterns in sequential data.

Based on the challenges faced in SER research for the Indonesian language, this study aims to address the gap by providing a comprehensive comparison of speech emotion recognition systems for determining emotional states. It evaluated the consistency and reliability of emotion labeling using Cohen's kappa, applied several feature extraction approaches including mel frequency cepstral coefficient (MFCC), zero crossing rate (ZCR), root mean square energy (RMSE), pitch, and spectral centroid, and combined these features with classification techniques that used CNN and LSTM. The structure of this paper is as follows: The study chronology, as well as the research design, methodology, dataset collection, feature extraction strategies, and classification algorithms, are covered in section 2. The research results appear in section 3 along with a comprehensive discussion, while section 4 provides the conclusion.

2. METHOD

In this research, the process of recognizing emotions in speech was organized into three main stages: data collection, feature extraction, and classification. During data collection stage, a dataset of speech samples representing various emotional states was gathered, and inter-rater reliability was employed to ensure consistent emotion labeling across different evaluators. Once the dataset was prepared, feature extraction was carried out to identify and process key characteristics of a speech signal. These extracted features were then used in the classification stage to accurately categorize the emotional states conveyed in the speech.

2.1. Data collection

The audio dataset for this research consists of speech recordings in Indonesian. The digital audio data was stored in WAV file format. The dataset includes recordings from 10 male and 10 female participants, aged 20 to 22 years. Each audio recording lasts between one to three seconds, with each participant contributing four recordings per emotion. Not all recordings, however, were suitable or usable due to factors such as poor audio quality or inconsistency in the emotional expression. A total of 50 audio files were used to represent four emotional expressions (angry, happy, neutral, and sad [12]), resulting in approximately 200 audio files in total. The dataset contained recordings with sampling rates that varied from 44.1 to 48 kHz. To ensure consistency for audio analysis, all files were resampled to 48 kHz, preserving high audio quality.

To evaluate the consistency and reliability of emotion labeling, Cohen's Kappa analysis was conducted [29]. This statistical method provides deeper insight into the agreement levels between annotators, using a scale from -1 to 1. A value of -1 represents complete disagreement, 0 indicates random agreement, and 1 reflects perfect agreement [30]. Table 1 contains Cohen's Kappa values and associated interpretations.

Table 1. Interpretation of Cohen's Kappa

Cohen's Kappa Statistic	Strength of agreement
< 0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

2.2. Feature extraction

One of the most important steps in processing speech data for emotion classification is feature extraction. This process involves transforming raw audio signals into relevant features for analysis. The techniques employed in this study were chosen to collect both temporal and spectral characteristics of speech. The techniques used in this study for extracting features from speech include MFCC, ZCR, RMSE, pitch, and spectral centroid. The features were extracted and calculated individually from each audio.

2.2.1. Mel frequency cepstral coefficient

The first feature extraction method employed was MFCC. MFCC is inspired by the way the human ear processes sound [31], [32]. These coefficients focus on the most important aspects of sound, such as the shape of vocal formants and other characteristics, which are essential for tasks like emotion recognition and speech analysis [33]. By emphasizing frequencies that are most important for how humans hear, MFCCs provide a clear representation of speech signals. Figure 1 illustrates the MFCC feature extraction process.

The extraction of MFCC features from speech data began with pre-emphasis, which boosted the higher frequencies to enhance clarity. Next, the audio signal was segmented into small frames of 25ms, with a 50% overlap. Each frame was then processed using a Hamming window to minimize edge effects before undergoing fast Fourier transform (FFT), which transformed the audio data from the time domain into the frequency domain using an NFFT size of 512. After that, a Mel-filter bank was applied, consisting of 40 filters spaced according to the Mel scale to mimic the human ear's frequency response. To manage the wide range of values, log compression was applied, compressing the values between 0 and 1. In the final step, the discrete cosine transform (DCT) was used on the log-compressed signal to derived MFCCs.

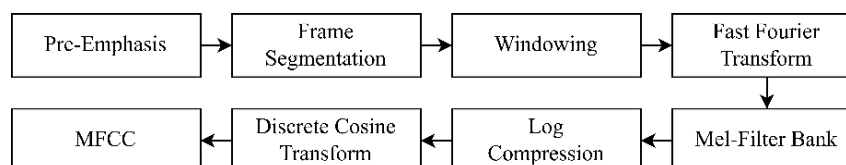


Figure 1. MFCC flowchart

2.2.2. Zero crossing rate

The second feature employed was the ZCR, which was calculated separately. ZCR was derived by assessing the frequency of zero-crossings in the signal across a frame. The process involved counting each

instance where the audio signal shifts from above zero to below zero or the reverse within a frame. ZCR is defined as shown in (1):

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

where N represents the total sample count within the frame, and $1(x[n] \cdot x[n-1] < 0)$ represents a function that outputs 1 when there is a sign change between $x[n]$ and $x[n-1]$, and 0 otherwise.

ZCR reflected the rate of change in the signal, which can be indicative of emotional states. Higher ZCR values are associated with more tense or agitated emotional states, such as angry or happy, where rapid changes in pitch or tone occur. Conversely, lower ZCR values indicate calmer emotions, such as neutral or sad, where the speech is more steady and less variable [34].

2.2.3. Root mean square energy

The third feature employed was the RMSE, or the root mean square value of a signal, which was derived by computing the square root of the mean value of the squared samples. For each sample $x[n]$ in the audio signal x , the square was calculated as: $x[n]^2$. The average of all resulting squared values was then calculated, and the square root of this average was used to determine the RMSE, as shown in (2):

$$E_{RMS} = \sqrt{\frac{1}{N} \sum_{n=1}^N x[n]^2} \quad (2)$$

where $x[n]$ is the signal value at index n , and N is the total number of signal samples.

RMSE reflected the intensity or volume of the speech signal. Emotions such as angry or happy involved higher energy levels due to louder and more forceful speech, resulting in higher RMSE values. Conversely, emotions like sad were expressed with softer, lower-energy speech, leading to lower RMSE values.

2.2.4. Pitch

The fourth feature analyzed was pitch. Pitch estimation from an audio signal involves several key steps to accurately determine the fundamental frequency or pitch. This process includes spectral analysis using techniques such as the fast Fourier transform (FFT)-based methods like autocorrelation or cepstral analysis. Pitch was calculated as shown in (3):

$$Pitch = \frac{\text{Sampling rate}}{\text{Index of Fundamental Frequency}} \quad (3)$$

Emotional states expressed through variations in the pitch of the voice. Emotions such as angry or happy tended to produce higher pitch variations, where the voice reached elevated frequencies, adding an energetic or intense quality to the speech. In contrast, sad or neutral involved a lower, more stable pitch, conveying a calmer or more subdued tone and signaling reduced emotional arousal.

2.2.5. Spectral centroid

The fifth feature calculated was the spectral centroid, which represents the center of gravity of the audio signal's frequency spectrum, providing an average frequency weighted by the amplitude of each spectral component. It is commonly used to describe how energy is distributed across the frequency range, offering insight into the brightness or sharpness of a sound. Spectral centroid was calculated using (4):

$$\text{Spectral Centroid} = \frac{\sum_{k=0}^{N-1} f(k) \cdot |X(k)|}{\sum_{k=0}^{N-1} |X(k)|} \quad (4)$$

where $f(k)$ is the frequency at index k , and $|X(k)|$ is the magnitude of the spectrum at index k .

Spectral centroid distinguished emotional states in speech by reflecting the brightness or sharpness of the voice. Higher spectral centroid values, linked to angry or happy, indicated energy concentrated in higher frequencies. Meanwhile, lower values, associated with neutral or sad, suggested a softer and more subdued tone.

Once each feature was extracted from each audio, a feature vector was formed to represent the key acoustic characteristics of the sound. This vector captured the most significant characteristics of the audio, which were essential in distinguishing various emotional states. These values were then used as input in the

classification process, where they helped the model recognize and distinguish between various types of emotional expressions.

2.3. Classification

The extracted features were used as inputs for emotion recognition through various classification techniques. In this study, both CNN and CNN-LSTM models were applied to compare in emotion recognition. These models were assessed to measure their accuracy in emotion classification using features extracted from the speech data. The experiments were conducted with data divided into 75% for training, 20% for testing, and 5% for validation. The models were implemented and tested in Google Colab.

2.3.1. Convolutional neural networks

The CNN model used was a 1D CNN designed to classify input data with 1D dimensions, such as time series or sensor data, where the order or relative position of the data is important. This model processed the input data through multiple convolutional layers, extracting spatial features that helped in emotion classification based on the speech data. Figure 2 depicts the architecture of a one-dimensional CNN model.

The 1D CNN architecture began with an input layer of size (23, 32), followed by several convolutional layers with filters of 32, 64, and 128, each accompanied by batch normalization, activation functions, and pooling layers to reduce data dimensionality. A dropout layer was then applied to prevent overfitting. The output from the last convolutional layer was flattened and passed to dense layers in order to get the final output, which was used to categorize the four emotions.

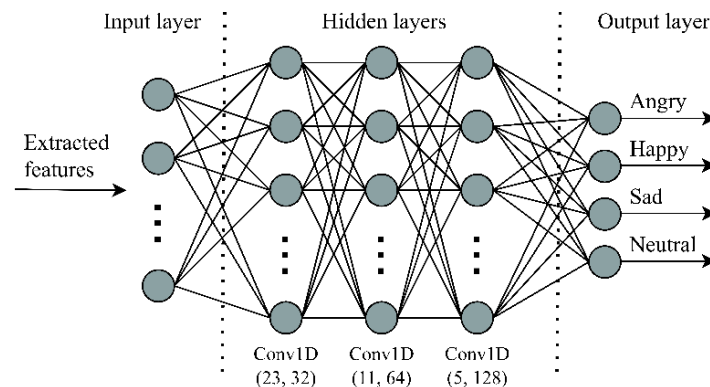


Figure 2. CNN1D architecture

2.3.2. Convolutional neural networks-long short-term memory

The CNN model followed by a LSTM network, often referred to as a CNN-LSTM model, is typically used for processing high-dimensional data such as audio or video. In this model, CNN retrieved spatial characteristics from the input, and LSTM captured the temporal dependencies among the derived features. Figure 3 depicts the architecture of the CNN-LSTM model.

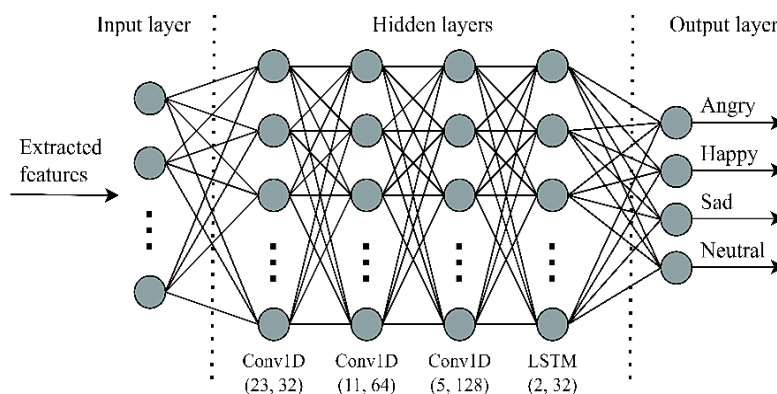


Figure 3. CNN-LSTM architecture

The CNN-LSTM architecture was similar to the 1D CNN architecture. In this model, the output from the convolutional layers was sent into an LSTM layer, which captured long-term dependencies in the data. The output from the LSTM layer was flattened and transferred to a dense layer that had four neurons, each of which represented one of the four emotional categories. This final dense layer served as the output, providing the predicted emotion based on the extracted features.

3. RESULTS AND DISCUSSION

This section provides the essential results of the research and a discussion on their significance. It explores aspects such as inter-rater reliability and compares the feature extraction methods and classification techniques used. These results provide insight into the effectiveness of emotion recognition from speech and highlight important trends observed throughout the analysis.

3.1. Inter-rater reliability results

The study found that the annotators had the highest agreement when labeling segments with the emotion "angry," achieving a score of 0.83. This suggests that "angry" was a distinct and easily recognizable emotion, leading to more consistent labeling among the annotators. The emotions "happy" and "sad" had agreement levels of 0.78 and 0.74, respectively. In contrast, "neutral" had the lowest agreement level, with a score of 0.72. This indicated that the absence of emotion or a neutral state is more subjective and harder to label consistently. The ambiguity and subtlety in neutral expressions likely contributed to this lower agreement level. The overall Cohen's Kappa results are illustrated in Figure 4.

The overall agreement across all emotions in the corpus was 0.69, which fell into the "substantial" agreement category. This overall Kappa value highlighted the variability and subjectivity in how the two annotators perceived and labeled emotions. The results of the dataset calculations using IBM SPSS software were summarized in Table 2.

The Kappa value of 0.698 indicated a substantial level of agreement between the annotators, suggesting that they often labeled emotions consistently. The T-value of 17.038 and a significance level of less than 0.001 confirmed that the findings were statistically significant, meaning that the observed results were unlikely to have occurred by chance. This indicated the reliability of the measurement process. However, these results also highlighted that while annotators generally agree, certain emotions led to inconsistencies in labeling.

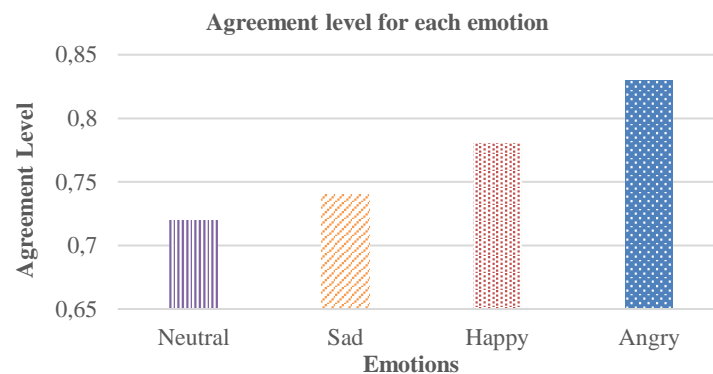


Figure 4. Cohen's Kappa results

Table 2. Cohen's Kappa results from IBM SPSS

Value	Asymptotic standard error ^a	Approximate T ^b	Approximate significance
0.698	0.039	17.038	<0.001

3.2. Feature extraction and classification comparison

Experiments were conducted using CNN and CNN-LSTM methods with input data derived from feature extraction of speech signals. The features extracted from the speech signals included MFCC, ZCR, RMSE, pitch, and spectral centroid. These features were analyzed to assess its impact in improving emotion classification accuracy.

3.2.1. Convolutional neural networks

In the CNN method, various feature combinations were tested to achieve the best results in emotion classification. These results highlighted the significance of each feature in helping the model to differentiate emotional states. Table 3 presents the accuracy of testing using CNN with different feature combinations.

Table 3. Accuracy comparison using CNN

Features	Accuracy
MFCC	81%
MFCC + ZCR	81%
MFCC + ZCR + RMSE	83%
MFCC + ZCR + RMSE + Pitch	88%
MFCC + ZCR + RMSE + Pitch + Spectral Centroid	85%

The use of MFCC alone resulted in an accuracy of 81%. MFCC is effective in capturing important spectral information, however using MFCC alone may not fully capture the temporal dimensions in speech that correspond to emotional states. Adding the ZCR to MFCC did not significantly improve accuracy. ZCR determines how frequently the signal crosses the zero-amplitude line within a specific time frame, but its contribution to emotion classification seemed less significant compared to other features. When RMSE was added to the combination of MFCC and ZCR, accuracy increased to 83%. RMSE, which measures the energy of the speech signal, enriches the representation of the temporal and strength aspects of the signal that are relevant for emotion detection. The increase in accuracy suggested that signal energy information played an important role in differentiating emotional expressions. Adding pitch to the combination of MFCC, ZCR, and RMSE led to a significant increase in accuracy to 88%. Pitch provides information about voice intonation, which is crucial in emotion recognition because variations in intonation can reflect deep emotional changes. The substantial contribution of pitch underscored the importance of intonation in distinguishing emotional expressions. However, adding Spectral Centroid to the combination of MFCC, ZCR, RMSE, and pitch slightly decreased accuracy to 85%. Spectral Centroid, which describes the center of mass of the spectral signal, did not seem to provide significant additional value in the context of emotion classification, or it might even complicate the model without adding informative value.

3.2.2. Convolutional neural networks-long short-term memory

The CNN-LSTM method also demonstrated strong performance in emotion classification. By combining the feature extraction capabilities with the sequential modeling power of LSTM, this approach captured both the relevant features from the speech signal and the sequential dependencies in the data. Table 4 shows the accuracy of testing using CNN-LSTM with various feature combinations.

Table 4. Accuracy comparison using CNN-LSTM

Features	Accuracy
MFCC	81%
MFCC + ZCR	77%
MFCC + ZCR + RMSE	88%
MFCC + ZCR + RMSE + Pitch	83%
MFCC + ZCR + RMSE + Pitch + Spectral Centroid	83%

Using MFCC alone resulted in an accuracy of 81%. MFCC is known to be effective in extracting spectral information from audio signals, but this accuracy suggested that the information obtained from MFCC alone still had limitations in fully detecting emotional variations. Adding the ZCR to MFCC actually reduced accuracy to 77%. This reduction might have been due to ZCR introducing noise or less relevant information, thereby disrupting the model's ability to classify emotions accurately. However, when RMSE was added to the combination of MFCC and ZCR, accuracy significantly increased to 88%. RMSE provides additional information about the intensity of the speech signal, which is crucial for distinguishing emotions. This increase in accuracy indicated that RMSE added crucial informative value for emotion detection. Adding Pitch to the combination of MFCC, ZCR, and RMSE increased accuracy to 83%. Although Pitch should provide additional information about the fundamental frequency of the voice relevant for emotion classification, this increase in accuracy was not as significant as in the previous combination. This might have been because the information provided by Pitch did not add enough value to the model or there was redundancy with existing features. Adding spectral centroid to the combination of MFCC, ZCR, RMSE, and Pitch did not further increase accuracy, which remained at 83%. Spectral centroid, which described the center

of mass of the sound spectrum, did not seem to provide sufficiently differentiating information compared to the other features or might have had an excessive overlap of information.

3.2.3. Comparison

The testing results for various feature extractions and classification methods show that RMSE had a significant impact in improving accuracy for both the CNN and CNN-LSTM models. In the CNN model, the combination of four features: MFCC + ZCR + RMSE + Pitch led to the highest accuracy of 88%. Similarly, the CNN-LSTM model reached the same accuracy with just three features: MFCC, ZCR, and RMSE. These results emphasized the significance of selecting the right features providing the most valuable contributions, while also highlighted an important trade-off between computational efficiency, as fewer features reduce the computational cost of feature extraction.

4. CONCLUSION

This study explored methods to enhance emotion recognition from Indonesian speech using feature extraction techniques and machine learning classification models. The experiments were conducted on an Indonesian language dataset consisting of 200 samples. To assess inter-rater reliability, Cohen's kappa analysis was conducted, which revealed a substantial agreement level ($\kappa = 0.698$) between annotators, highlighting the consistency of emotion labeling. The classification experiments compared CNN and CNN-LSTM models. Both the CNN model, which used four features (MFCC, ZCR, RMSE, and Pitch), and the CNN-LSTM model, which used three features (MFCC, ZCR, and RMSE), achieved an emotion classification accuracy of approximately 88%. The difference in the number of features suggests that while the CNN model involved more computational tasks due to the additional feature, the CNN-LSTM model managed to achieve similar performance with fewer features, potentially offering a more efficient approach.

Overall, the findings demonstrate that incorporating diverse feature extraction techniques can enhance emotion recognition performance, particularly in Indonesian SER. However, careful consideration is needed to balance computational efficiency and feature complexity, as adding more features can improve accuracy but may also increase computational cost. Future research could explore the use of advanced optimization techniques or feature selection methods to further refine model performance while minimizing computational overhead.

ACKNOWLEDGMENTS

The authors would like to thank PENS management for all support in the form of laboratory facilities and all the equipment provided, so that we can carry out this research well.

FUNDING INFORMATION

This research has been supported by the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia with the scheme of *Penelitian Tesis Magister* for the fiscal year 2024, with the contract number of 524/PL14/PT.01.05/III/2024.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Izza Nur Afifah	✓	✓	✓	✓	✓	✓		✓	✓		✓		✓	
Tri Budi Santoso		✓				✓		✓	✓	✓		✓		✓
Titon Dutono			✓		✓		✓		✓					

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding authors, INA and TBS, upon reasonable request.




REFERENCES

- [1] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, Jan. 2020, doi: 10.1016/j.specom.2019.12.001.
- [2] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47795–47814, 2021, doi: 10.1109/access.2021.3068045.
- [3] M. Bojanić, V. Delić, and A. Karpov, "Call redistribution for a call center based on speech emotion recognition," *Applied Sciences*, vol. 10, no. 13, p. 4653, Jul. 2020, doi: 10.3390/app10134653.
- [4] V. M. Koti, K. Murthy, M. Suganya, M. S. Sarma, G. V. S. S. Seshu Kumar, and B. N., "Speech emotion recognition using extreme machine learning," *EAI Endorsed Transactions on Internet of Things*, vol. 10, Nov. 2023, doi: 10.4108/eetiot.4485.
- [5] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis, "Speech technology for healthcare: opportunities, challenges, and state of the art," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 342–356, 2021, doi: 10.1109/rbme.2020.3006860.
- [6] H. H. Mustafa, N. R. Darwish, and H. A. Hefny, "Automatic speech emotion recognition: a systematic literature review," *International Journal of Speech Technology*, vol. 27, no. 1, pp. 267–285, Mar. 2024, doi: 10.1007/s10772-024-10096-7.
- [7] S. Langari, H. Marvi, and M. Zahedi, "Efficient speech emotion recognition using modified feature extraction," *Informatics in Medicine Unlocked*, vol. 20, p. 100424, 2020, doi: 10.1016/j.imu.2020.100424.
- [8] A. Hashem, M. Arif, and M. Alghamdi, "Speech emotion recognition approaches: A systematic review," *Speech Communication*, vol. 154, p. 102974, Oct. 2023, doi: 10.1016/j.specom.2023.102974.
- [9] M. Liu, "English speech emotion recognition method based on speech recognition," *International Journal of Speech Technology*, vol. 25, no. 2, pp. 391–398, Feb. 2022, doi: 10.1007/s10772-021-09955-4.
- [10] M. H. Pham, F. M. Noori, and J. Torresen, "Emotion recognition using speech data with convolutional neural network," in *2021 IEEE 2nd International Conference on Signal, Control and Communication (SCC)*, Dec. 2021, pp. 182–187, doi: 10.1109/scc53769.2021.9768372.
- [11] D. Issa, M. Fatih Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, May 2020, doi: 10.1016/j.bspc.2020.101894.
- [12] F. Fahmi, M. A. Jiwanggi, and M. Adriani, "Speech-emotion detection in an Indonesian movie," in *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, 2020, pp. 185–193.
- [13] P. Kurniawati, D. P. Lestari, and M. L. Khodra, "Speech emotion recognition from Indonesian spoken language using acoustic and lexical features," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, Nov. 2017, pp. 1–7, doi: 10.1109/icsda.2017.8384467.
- [14] N. B. Wunarno and Y. E. Soelistio, "Towards Indonesian speech-emotion automatic recognition (I-SpEAR)," in *2017 4th International Conference on New Media Studies (CONMEDIA)*, Nov. 2017, pp. 98–101, doi: 10.1109/conmedia.2017.8266038.
- [15] O. U. Kumala and A. Zahra, "Indonesian speech emotion recognition using cross-corpus method with the combination of MFCC and Teager energy features," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 4, 2021, doi: 10.14569/ijacsa.2021.0120422.
- [16] T. Liu and X. Yuan, "Paralinguistic and spectral feature extraction for speech emotion classification using machine learning techniques," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, May 2023, doi: 10.1186/s13636-023-00290-x.
- [17] A. Aggarwal *et al.*, "Two-way feature extraction for speech emotion recognition using deep learning," *Sensors*, vol. 22, no. 6, p. 2378, Mar. 2022, doi: 10.3390/s22062378.
- [18] S. Sekkate, M. Khalil, and A. Adib, "A statistical feature extraction for deep speech emotion recognition in a bilingual scenario," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11443–11460, Oct. 2022, doi: 10.1007/s11042-022-14051-z.
- [19] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, "Deep learning approaches for speech emotion recognition: state of the art and research challenges," *Multimedia Tools and Applications*, vol. 80, no. 16, pp. 23745–23812, Jan. 2021, doi: 10.1007/s11042-020-09874-7.
- [20] S. Jain and B. Kishore, "Comparative study of voice print Based acoustic features: MFCC and LPCC," *International Journal of Advanced engineering, Management and Science*, vol. 3, no. 4, pp. 313–315, 2017, doi: 10.24001/ijaems.3.4.5.
- [21] S. Joo, J. Choi, N. Kim, and M. C. Lee, "Zero-crossing rate method as an efficient tool for combustion instability diagnosis," *Experimental Thermal and Fluid Science*, vol. 123, p. 110340, May 2021, doi: 10.1016/j.expthermflusci.2020.110340.
- [22] S. Cunningham, H. Ridley, J. Weinel, and R. Picking, "Supervised machine learning for audio emotion recognition: Enhancing film sound design using audio features, regression models and artificial neural networks," *Personal and Ubiquitous Computing*, vol. 25, no. 4, pp. 637–650, Apr. 2020, doi: 10.1007/s00779-020-01389-0.
- [23] R.-A. Knight and J. Setter, *The Cambridge handbook of phonetics*. Cambridge University Press, 2021, doi: 10.1017/9781108644198.
- [24] J. M. K. Kua, T. Thiruvanan, M. Nosratighods, E. Ambikairajah, and J. Epps, "Investigation of spectral centroid magnitude and frequency for speaker recognition," in *Odyssey 2010: Speaker and Language Recognition Workshop*, 2010, pp. 34–39.
- [25] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *Proceedings of the 22nd ACM international conference on Multimedia*, Nov. 2014, pp. 801–804, doi: 10.1145/2647868.2654984.
- [26] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Interspeech 2015*, Sep. 2015, doi: 10.21437/interspeech.2015-336.




- [27] T. Anvarjon, Mustaqeem, and S. Kwon, "Deep-Net: a lightweight CNN-based speech emotion recognition system using deep frequency features," *Sensors*, vol. 20, no. 18, p. 5212, Sep. 2020, doi: 10.3390/s20185212.
- [28] A. Aftab, A. Morsali, S. Ghaemmaghami, and B. Champagne, "LIGHT-SERNET: a lightweight fully convolutional neural network for speech emotion recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 6912–6916, doi: 10.1109/icassp43922.2022.9746679.
- [29] C. Vogel and K. Ahmad, "Agreement and disagreement between major emotion recognition systems," *Knowledge-Based Systems*, vol. 276, p. 110759, Sep. 2023, doi: 10.1016/j.knosys.2023.110759.
- [30] M. L. McHugh, "Interrater reliability: the Kappa statistic," *Biochemia Medica*, pp. 276–282, 2012, doi: 10.11613/bm.2012.031.
- [31] D. M. Nogueira, C. A. Ferreira, E. F. Gomes, and A. M. Jorge, "Classifying heart sounds using images of motifs, MFCC and temporal features," *Journal of Medical Systems*, vol. 43, no. 6, May 2019, doi: 10.1007/s10916-019-1286-5.
- [32] S. D. Waghmare, R. R. Deshmukh, P. P. Shrishrimal, V. B. Waghmare, and G. B. Janvale, "Stuttered isolated spoken Marathi speech recognition by using MFCC and LPC," *International Journal of Innovations in Engineering and Technology*, vol. 8, no. 3, 2017, doi: 10.21172/ijiet.83.018.
- [33] M. M. Rezapour Mashhadi and K. Osei-Bonsu, "Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest," *PLOS ONE*, vol. 18, no. 11, p. e0291500, Nov. 2023, doi: 10.1371/journal.pone.0291500.
- [34] H. K. Palo, "The effect of age, gender, and arousal level on categorizing human affective states," in *Emotion and Information Processing*, Springer International Publishing, 2020, pp. 97–124, doi: 10.1007/978-3-030-48849-9_7.

BIOGRAPHIES OF AUTHORS






Izza Nur Afifah    received her Bachelor of Applied Science degree in telecommunication engineering from Politeknik Elektronika Negeri Surabaya, Indonesia, in 2022. Currently, she is pursuing a Master of Applied Science degree in computer and informatics engineering at the same institution. Her research interests during her study include speech emotion recognition and machine learning. She can be contacted at email: izzanurafifah@gmail.com.



Tri Budi Santoso    received his B.Eng. degree in engineering physics, his M.T. degree in electrical engineering, and his Dr. degree from the Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia, in 1994, 1999, and 2016, respectively, with research focused on signal processing. He has been with the Department of Electrical Engineering at the Electronic Engineering Polytechnic Institute of Surabaya (PENS) since 1995. His research interests include telecommunications and acoustic signal processing. He can be contacted at email: tribudi@pens.ac.id.



Titon Dutono    was born in Surabaya, Indonesia, in 1960. He received a B.S. degree in telecommunication engineering from Sepuluh Nopember Institute of Technology, Surabaya, Indonesia, in 1985, and both his Master's and Doctor of Engineering degrees in electrical engineering and computer science from Kumamoto University, Kumamoto, Japan, in 1994 and 1997, respectively. From 2002 to 2008, he served as the Principal of the Electronics Engineering Polytechnic Institute of Surabaya (PENS). From 2008 to 2016, he was appointed Deputy Director-General for Spectrum Policy and Planning at the Ministry of Communication and Information Technology, Republic of Indonesia. He was also in charge of leading the Indonesian delegation during ITU regulatory meetings in Geneva and other venues. Since 2017, he has returned to campus as an associate professor in the Electrical Engineering Department of EEPIS. His research interests include signal processing, radio communication, telecommunication regulation, and teaching methodology. He can be contacted at email: titon@pens.ac.id.