The Indonesian Journal of Computer Science
www.ijcs.net
Volume 13, Issue 6, December 2024
https://doi.org/10.33022/ijcs.v13i6.4442

Speech Emotional Recognition of Telephone Conversation by Using Deep Learning

Izza Nur Afifah[1], Titon Dutono[2], Tri Budi Santoso[3]
izzanurafifah@gmail.com[1], titon@pens.ac.id[2], tribudi@pen.ac.id[3]
[1,2] Electrical Department, Politeknik Elektronika Negeri Surabaya
[3] Multimedia Creative Department, Politeknik Elektronika Negeri Surabaya

Article Information

Received : 15 Oct 2024
Revised : 5 Nov 2024
Accepted : 5 Dec 2024

Keywords

Speech emotional recognition; MFCC; CNN; RNN.

Abstract

In this research, we have compared two clustering algorithms, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) to support the SER model for voice communication. We proposed an approach to speech emotion recognition in telephone conversations using a combination of Mel-frequency cepstral coefficients (MFCC) for audio feature extraction. Only the zeroth coefficient (energy) of MFCC will be used, as energy can provide a good representation for sound. The extracted results are then classified using CNN and RNN. The CNN algorithm consistently achieved a higher accuracy than RNN with the values of 0.93 at epoch=50, 0.93 at epoch=100, 0.90 at epoch=150, and 0.93 at epoch=200. But the RNN algorithm has a faster training times than CNN. For optimizer test, Adam optimizer performs well for both models with respective accuracy values of 0.93 and 0.94, outperforming other optimizers for accuracy.

A. Introduction

One of the developments in Human Computer Interaction (HCI) technology in human-robot communication systems has gradually shifted from command communication to emotional communication. It involves various difficulties and challenges in the interaction process. In this research the application of gesture recognition and speech recognition in HCI technology has also been explained. The results have shown that a combination of intelligent HCI with deep learning is one of the hopes in realizing the applications of gesture recognition, speech recognition, emotion recognition, and intelligent robot direction, [1- 3].

One commonly used approach in emotion recognition is the analysis of speech features extracted from speech signals. In this research, Mel-Frequency Cepstral Coefficients (MFCC) are utilized as the chosen method for feature extraction. MFCC has proven to be effective in capturing essential voice characteristics [4], including the emotions expressed in telephone conversations.

The extracted features will be utilized to train a classification model for emotion recognition in telephone conversations. Efficient classifier techniques like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) can be employed to classify the emotions contained.

A study has been proposed the using Mel-Frequency Cepstral Coefficients (MFCC) as the main feature in emotion recognition. The application of MFCC to voice data to identify key emotions, such as Anger, Fear, Joy, Sadness, Boredom, and Neutral [5]. The result showed that MFCC can generate useful representations of sound signals related to human emotions.

A Convolutional Neural Network (CNN) has become a popular approach in speech pattern recognition, including emotion recognition. This study proposed the using of CNN architecture, as well as other techniques in emotion recognition [6]. They demonstrated that CNN able to yield good results in classifying emotions from human speech signals.

Research of the Deep RNN architecture, which combines RNN with many layers (deep layers), to train speech recognition models has been carried out. They compared the performance of their Deep RNN model with several methods including Hidden Markov Models (HMM) and Deep Belief Networks (DBN). The results showed that the Deep RNN model managed to achieve better performance in speech recognition compared to HMM and DBN in terms of recognition accuracy and training speed [7].

The objective of this research is to discover more effective algorithms that optimize features and enhance the performance of emotion recognition, ultimately contributing to the development of a sophisticated emotion recognition system for telephone conversations as a system that has been proposed by the previous researcher [8]. In this research the features extraction conducted by using MFCC and classification step supported with Convolutional Neural Network (CNN) and recurrent neural network (RNN) as the previous research [9-12]. There two algorithms are analysed and compare to find out which one is suitable for the application of speech emotion recognition conversation on the telephone communications.

B. Research Method

The speech emotional detection involves three main stages. The first stage is data collection, which is obtained from voice telephone. The second stage is pre-processing, which involves a series of steps to process the collected voice data. In this the feature of speech signal is extracted by using Mel-frequency Cepstral Coefficients (MFCC). After the features are extracted, there are two kinds of classification process, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) algorithms. Overall, the stages of system design as in Figure 1.

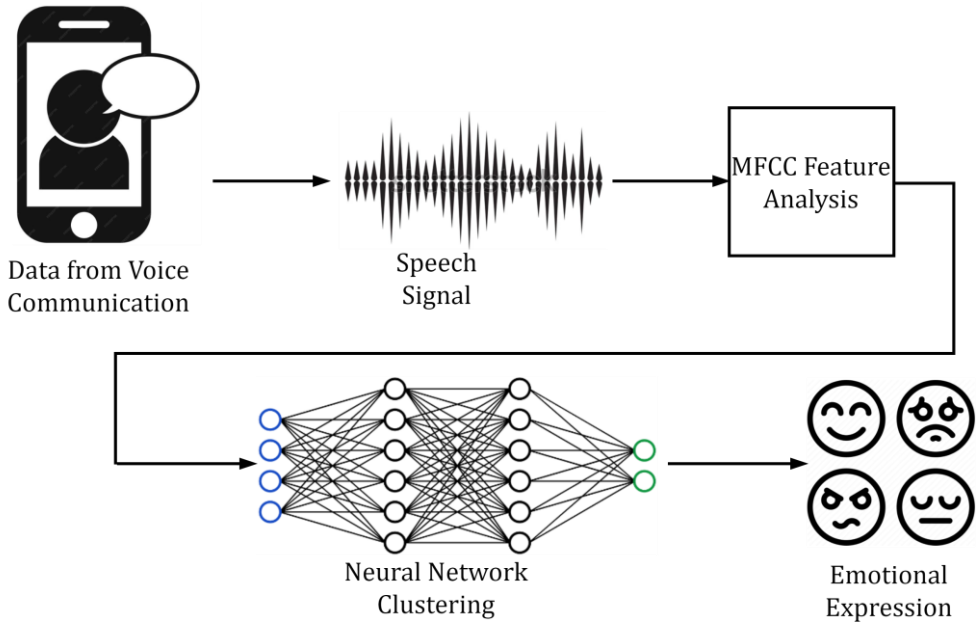


Figure1. System design

Data Collecting

This research utilizes voice data that has been previously recorded and stored in .wav format. Each file consists of samples of four different emotions (happy, sad, angry and neutral) with length of 1 ~ 3 seconds. Every emotion expression is represented by 50 sound samples, therefore the total sample data used in this study is 200 samples as in Table 1.

Table 1. The label of file expression

No	Emotion	*Size*
1	Angry	50
2	Hapy	50
3	Sad	50
4	Neutral	50

The datasets are organized into folders and appropriately labeled according to defined naming standards containing the name of the speaker, a numeric identifier, and the type of emotion expressed in the file as in Fig 2.

Audio data is divided during the classification process into an 80:20 ratio for training and testing purposes. Therefore, 80% of the data is used to train the

classification model, while the remaining 20% is set aside to evaluate its performance.

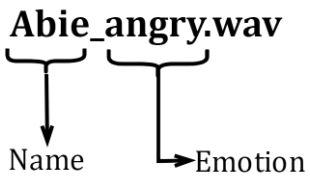


Figure 2. Audio files labeling.

Pre-processing Feature Extraction

The audio dataset contains recordings with varying sample rates, so resampling is necessary for consistency for further audio analysis. Resampling involves adjusting the sample rate of each audio file to make sure they all have the same rate. In this case, the sample rate ranges from 44.1 kHz to 48 kHz. To maintain high audio quality, all audio files are re-sampled to 48 kHz.

The feature extraction process uses Mel-frequency Cepstral Coefficients (MFCC). Process the speech signal processing to obtain the MFCC coefficient as shown in Figure 3.

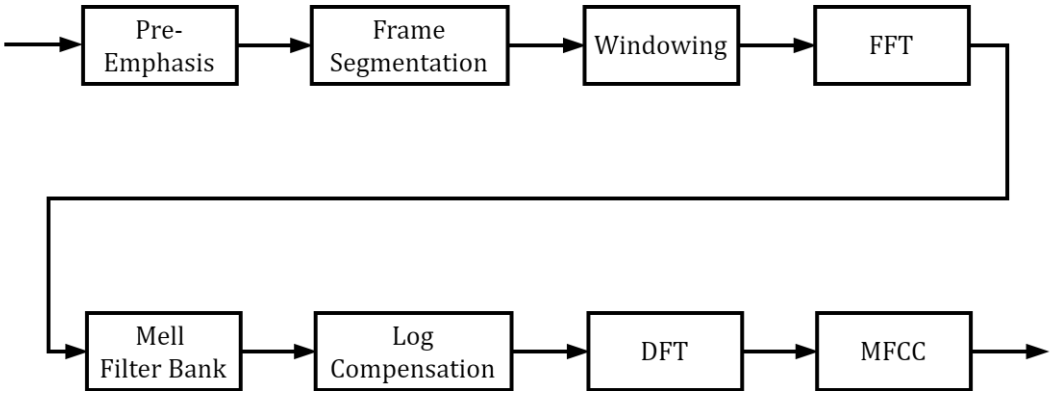


Figure 3. MFCC feature extraction

In MFCC feature extraction, a pre-emphasis filter with a constant α is set on 0.97. This technique increases the high-frequency component while slightly reducing the low-frequency intensity, effectively. The general formula of pre-emphasis is as follows:

$$y[n] = s[n] - \alpha s[n-1]; 0{,}9 \leq \alpha \leq 1.0 \qquad (1)$$

The audio signal is divided into small frames. The frame size truncation is defined as 25 milliseconds (ms) and the step between successive frames is 10ms. Segmented signals are processed using a technique called windowing, which is useful for reducing visible discontinuities at frame boundaries. Window Hamming was chosen because it is widely used in speech recognition applications, produces a relatively low sidelobe level, and also reduces the resulting noise level. The next step involves applying the Fourier transform to each window of the sound signal. The result of the Fourier transform is a spectrum with complex values that represent the frequency content of the sound signal. After that, Mel Filter-bank is

applied to produce a set of filters used in feature extraction to analyze speech signals. This set of filters represents the speech signal in the Mel frequency domain, which is commonly used in speech recognition and other audio processing tasks to describe the distribution of frequency energy with accuracy consistent with human perception. After applying a bank filter to the audio signal, the next step is to apply a Discrete Cosine Transform (DCT) to the output of the bank filter. The DCT process on the filter bank produces an MFCC matrix which contains the MFCC coefficients.

In this study, only the zero coefficient (energy) is used. The emphasis on the zero coefficient (energy) is based on the reason that energy information can provide a good representation for speech recognition and classification. Using a zero coefficient also helps reduce computational complexity and data dimensions in the classification stage.

Convolutional Neural Network

The input dimensions for the CNN model are adjusted to the size of the features used. The CNN model is built using Sequential from Keras, which consists of several interconnected layers. Because the number of existing datasets is 200 files, the CNN model used has a simple structure with the aim of avoiding overfitting. The appearance of the CNN model in this study is shown in Figure 4.

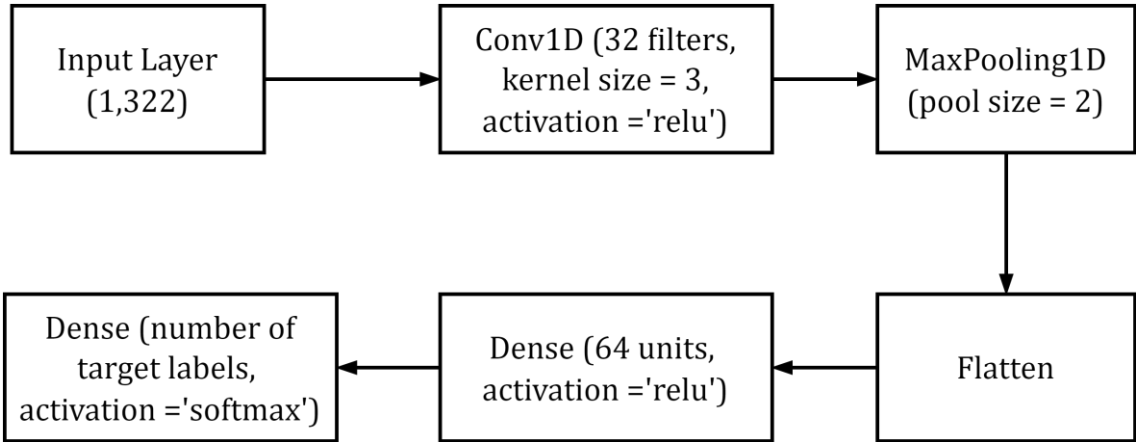


Figure 4. CNN model

First, there is a Conv1D layer with 32 filters and a kernel size of 3. This layer is responsible for extracting important features from the input data. Then, it is followed by a MaxPooling1D layer that takes the maximum value from each feature generated by the Conv1D layer. This helps in reducing the dimensionality of the data and selecting more important features.

There is a Flatten layer that made a flat the output from the previous layer into a one-dimensional vector. This allows the data to be used by the next Dense layer. The next layer in the model is a Dense layer with 64 neurons and the relu activation function. This layer processes the previously extracted features. In the last layer, there is a Dense layer with a few neurons matching the number of target categories in the problem at hand. The softmax activation function is used in this layer to generate a probability distribution for each target category.

After the model is built, the model is then compiled using categorical_crossentropy as a loss function, various optimizers, and accuracy metrics to evaluate model performance during the training process.

Recurrent Neural Network

The Recurrent Neural Network (RNN) model uses the Sequential API from Keras. The model consists of an LSTM layer with 128 units, a Dense layer with 64 units, and a Dense layer with 4 units using the softmax activation function. The number of units in the last Dense layer is adjusted to the number of classes in the label. The appearance of the RNN model in this study is shown in Figure 5.

After building the model, it is compiled by using the 'sparse_categorical_crossentropy' loss function and various optimizers available.

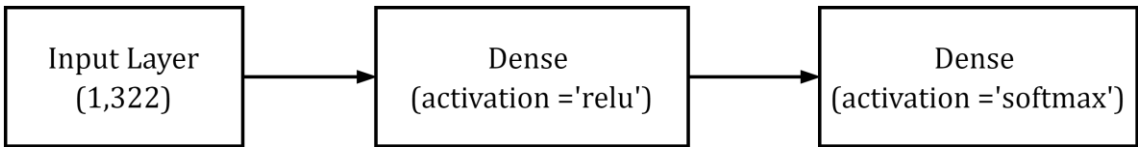


Figure 5. RNN model

C. Result and Discussion

Feature Comparison

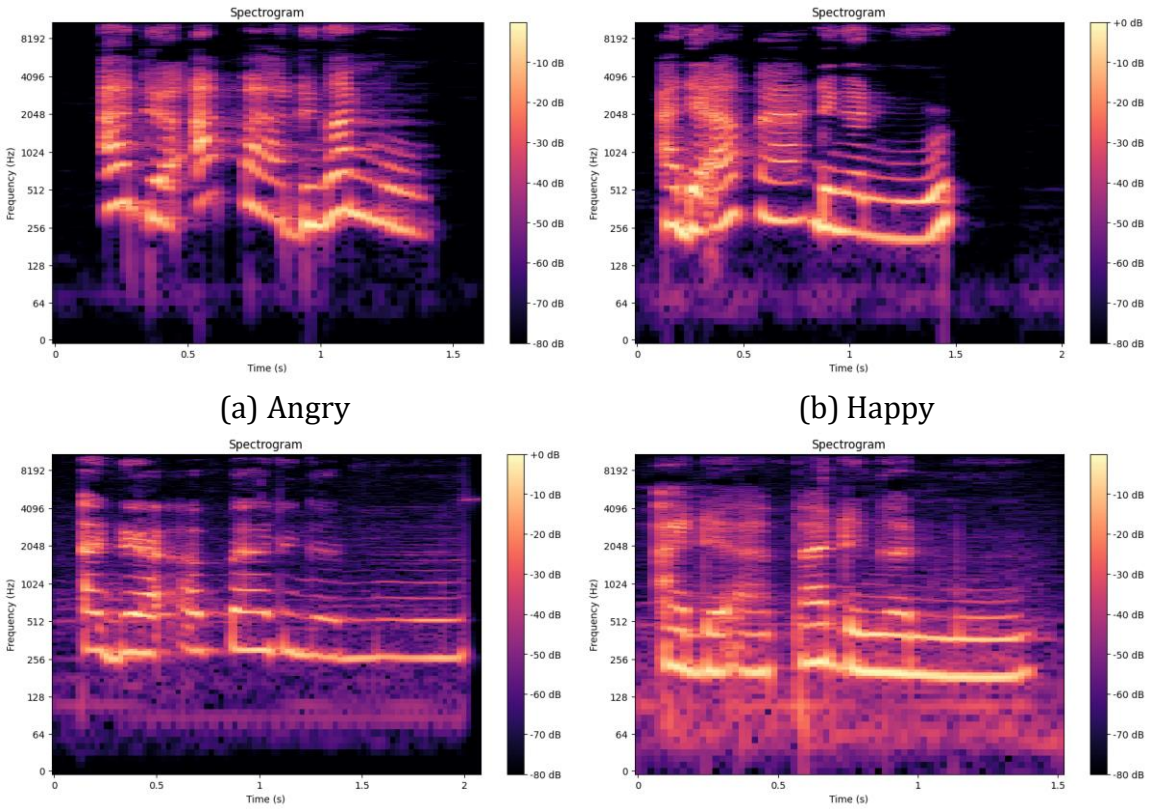
After a series of feature extraction processes, a spectrogram image of the speech signal is obtained for the 4 types of expressions that have been determined. In simple terms, for several speech signal samples, an output picture is obtained as shown in Figure 6.

The Angry expression (Figure 6a) shows that the energy comes from the frequency area between 256 ~ 4098 Hz. There is a rise and fall pattern in frequency as a function of time. This is understood as a natural behavior when someone is angry, showing a pounding heartbeat, causing energy fluctuation output.

For the Happy expression (Figure 6b) energy does not appear throughout the entire time, but when it appears it has a high energy. This is marked with bright color and is centered on several frequency values (256 Hz and 500 Hz) which also shows the dominance of the energy value. This represents when a speaker is showing a happy expression and can make an explosive expression with a certain pattern.

The Sad expression (Figure 6c) has a different spectrogram pattern. In this case the energy tends to weaken and the frequency components that appear are also relatively reduced. In terms of time, it looked that it tends to be fully charged, but with weak energy. This condition can be understood that when a speaker is in a sad condition, the sound produced tends to weaken and appears with relatively low energy.

In the Neutral expressions (Figure 6d) the speech signal tends to cover the entire time. In this case, the frequency side also shows that almost all frequency components are filled, but with relatively low energy. This is indicated by the red color which tends to be dark, according to the dB scale on the right of the image.



(a) Angry (b) Happy

(c) Sad (d) Neutral

Figure 6. Feature Comparison

The extraction process using MFCC basically has a shape similar to the pattern that appeared in the Spectrogram. However, in this case the graph that appears tends to have a resolution that is not fine, because it is a display of the output coefficient values with a frequency function that has been scaled with Mell-scale. The MFCC output value is then used as input for the clustering process using two different algorithms, namely CNN and RNN.

Accuracy and Time Computation

The purpose of accurate test is to get the more exploration to the model's ability to predict and classify sounds with different emotions. The architecture testing of accuracy and time parameters for each epoch value. The first test was carried out with the number of epoch parameters between 50-200. Therefore, the accuracy of the prediction results of emotion and time is obtained. Table II below shows a comparison of accuracy and computation time for each classification algorithm with various epoch values.

After the test conducted, the CNN algorithm indicated to achieve a higher level of accuracy than the RNN algorithm, but this algorithm needs more time to train the model compared to the RNN algorithm. Generally, the time required by both algorithms increases as the number of epochs increases. This suggests that the more epochs trained, the longer it would take.

Table 2. Comparison of Accuracy and Time Computation

Epoch	Algorithm	Accuracy	Time (second)
50	CNN	0.93	43
	RCC	0.89	27
100	CNN	0.93	54
	RCC	0.875	36
150	CNN	0.9	69
	RCC	0.89	53
200	CNN	0.93	90
	RCC	0.92	49

Confusion Matrix

The confusion matrix from the training results is shown in Figure 7a for CNN and 7b for RNN algorithms. The numbers in the diagonal column represent data detected as genuine emotions, while the other columns indicate detection errors.

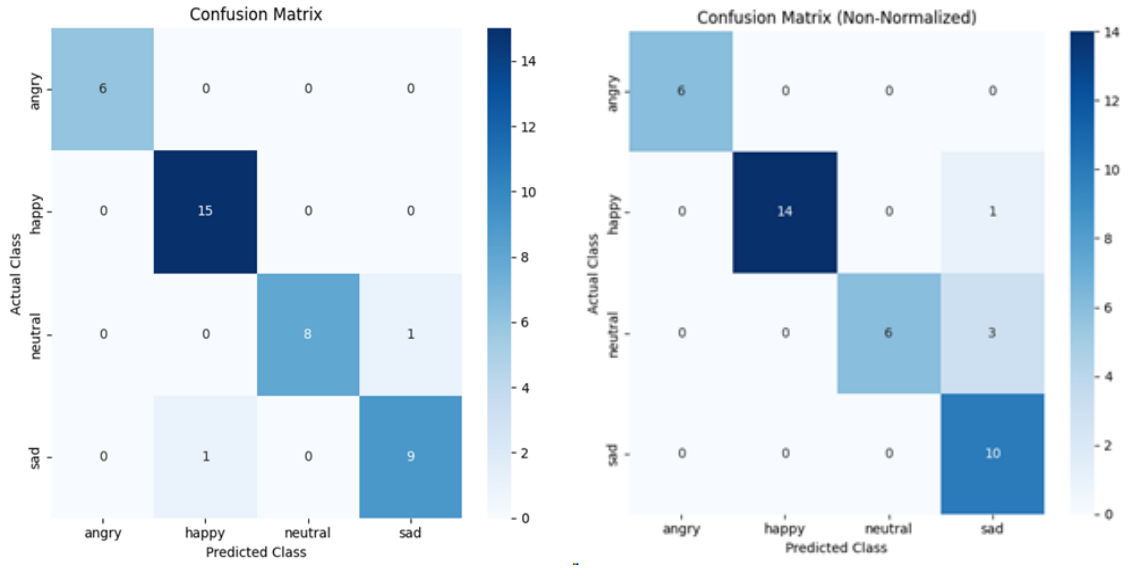
The experiment result for emotional expressions indicated that the CNN algorithm showed a good performance for Angry and Happy expression. There is no error happen along testing phase conducted. For the Sad and Neutral expressions this algorithm showed one error, and this is indicated that this algorithm is still good and achievable.

In the case Sad and Neutral expression, it conveyed in a flat tone of voice are often predicted as other expression. For Sad expression with a flat tone is predicted as Neutral. But a strange incident has appeared in one Neutral expression that was predicted to be Happy. In this case, a relatively flat tone pattern should be predictable as Sad. This phenomenon can occur due to the system's inability to accept a tone pattern that may appear beyond its prediction.

In experiments conducted by using the RNN algorithm showed a different behavior compared to CNN. In this case the system tends to display perfect performance in the expression of Angry and Sad emotions. In fact, these two emotional expressions have different physical behavior, where Angry tends to carry a fluctuating tone with high energy, while Sad carries a relatively flat tone with weak energy.

In the Happy emotion expression, the system with the RNN algorithm showed good performance, because only one error occurred. This happens when a Happy expression is predicted to be a Sad expression. In Neutral emotion expression, the system with the RNN algorithm does not show good performance. In this case, three errors have occurred, in which the Neutral expression is predicted as a Sad expression. Systems with the RNN algorithm show behavior patterns that attract attention. In this case, the prediction errors that occur are somewhat difficult to relate to the logic of existing physical patterns, in this case the pattern of vocal tones produced by each emotional expression.

Overall, after normalizing the confusion matrix, it was found that Angry, Happy and Sad emotions tend to be easier to recognize compared to Neutral emotions.



(a) CNN (b) RNN

Figure 7. Confusion Matrix Comparison

Optimizer Test

Subsequent tests were conducted on each optimizer. The tests are carried out with the number of epoch parameters being 200. And the learning rate used was 0.001 for sgd optimization. Table 3 below shows a comparison of the accuracy for each optimization. The optimizers tested were sgd, adam, rmsprop, and adagrad.

Table 3. Comparison of Optimizer Accuracy

Optimizer	Algorithm	Score
Sgd	CNN	0.93
	RCC	0.75
Adam	CNN	0.93
	RCC	0.94
RMSdrop	CNN	0.88
	RCC	0.89
adamgrad	CNN	0.82
	RCC	0.87

Analysis of the above table shows that the optimizer performance can vary depending on the type of model used. In this case, Adam gives better results on the CNN model, while SGD gives better results on the RNN model. Likewise, RMSprop gives better results on the CNN model, while Adagrad gives better results on the RNN model.

In the selection of optimizer, we have to consider the type of model used and to conduct experiments to find the most suitable optimizer for a particular model. Besides that, other parameters in the optimizer such as learning rate can be adjusted to achieve more optimal results.

D. Conclusion

SER has become an interesting topic that has been widely developed using deep learning techniques. In this research, we have compared two clustering algorithms, CNN and RRNN to support the speech emotion recognition (SER) model for voice communication.

Based on the analysis provided, we can draw the following conclusions that CNN consistently achieves higher accuracy compared to RNN across all epochs, with accuracy values of 0.93 at epoch=50, 0.93 at epoch=100, 0.90 at epoch=150, and 0.93 at epoch= 200. If accuracy is the primary concern, CNN is a better choice. On the other hand, RNN requires less time for training compared to CNN at each epoch. At epoch=200, with an accuracy almost close to CNN, RNN only takes 49s at an accuracy of 0.92 compared to CNN which takes 90.56s at an accuracy of 0.93. If reducing training time is a priority, RNN is more efficient.

The choice between CNN and RNN depends on the specific requirements of the application, considering the trade-off between accuracy and training time. Meanwhile, in the optimizer test, adam optimizer performs well on both CNN and RNN models with respective values of 0.93 and 0.94 compared to the other optimizers in this research.

E. Acknowledgment

We would like to thank to the Ministry of Education, Culture, Research and Technology for supporting this research through the 'Thesis Magister Research' funding scheme for Fiscal Year 2024, and to PENS Management for providing laboratory facilities so that this research can be carried out well.

F. References

[1] L. Zhihan, F. Poiesi, Q. Dong, J. Lloret, and H. Song, "Deep Learning for Intelligent Human–Computer Interaction", *MDPI, Applied Science*, 2022, Vol. 12, 11457.

[2] N. Gradega, C. Busso, E. Alvarado, H.F. Fernando, R. García, B. Yoma, and R. Mahu, "Speech Emotion Recognition in Real Static and Dynamic Human-Robot Interaction Scenarios", *ScienceDirect, Computer Speech and Language*, vol. 89. 2025.

[3] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human–computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005, doi: https://doi.org/10.1016/j.neunet.2005.03.006.

[4] N. Singh, R. A. Khan, and R. Shree, "MFCC and prosodic feature extraction techniques: a comparative study," *International Journal of Computer Applications*, vol. 54, no. 1, 2012.

[5] P. P. Dahake, K. Shaw and P. Malathi, "Speaker dependent speech emotion recognition using MFCC and Support Vector Machine," *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, Pune, India, 2016, pp. 1080-1084, doi: 10.1109/ICACDOT.2016.7877753.

[6] A. B. Abdul Qayyum, A. Arefeen and C. Shahnaz, "Convolutional Neural Network (CNN) Based Speech-Emotion Recognition," *2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)*, Dhaka, Bangladesh, 2019, pp. 122-125, doi: 10.1109/SPICSCON48833.2019.9065172.

[7] A. Graves, A. -r. Mohamed and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, 2013, pp. 6645-6649, doi: 10.1109/ICASSP.2013.6638947

[8] Q. Zhao, K. Chen, T. Li, Y. Yang, and X. Wang, "Detecting telecommunication fraud by understanding the contents of a call," *Cybersecurity*, vol. 1, no. 1, p. 8, 2018, doi: 10.1186/s42400-018-0008-5.

[9] Fahmi, M.A. Jiwanggi, and M. Adriani, M., "Speech-Emotion Detection in an Indonesian Movie", *Language Resources and Evaluation Conference (LREC 2020)*, Marseille, 11–16 May, 2020.

[10] R.Y. Rumagit, G. Alexander, I.F. Saputra, "Model Comparison in Speech Emotion Recognition for Indonesian Language", *5th International Conference on Computer Science and Computational Intelligence 2020*, ScienceDirect, Procedia Computer Science, Vol. 179, pp:789–797. 2021

[11] T. Dutono, F.M. Nuriyah, and T.B.Santoso, "Instrumental Music Emotion Recognition with MFCC and KNN Algorithm", *Indonesian Journal of Computer Science,* vol. 12, No. 1, pp. 64-72. 2023.

[12] S. Yusdiantoro, and T.B. Sasongko, "Implementasi Algoritma MFCC dan CNN dalam Klasifikasi Makna Tangisan Bayi", *Indonesian Journal of Computer Science,* vol. 12, No. 4, pp. 1957-1968. 2023.