Speech emotion recognition with optimized multi-feature stack using deep convolutional neural networks

Muhammad Farhan Fadhil, Amalia Zahra
Department of Computer Science, BINUS Graduate Program, Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Feb 18, 2023
Revised Apr 29, 2024
Accepted May 17, 2024

Keywords:

Convolutional neural network
Mel frequency cepstral coefficient
Multi-feature stack
Optimized deep convolutional neural network
Speech emotion recognition

ABSTRACT

The human emotion in communication plays a significant role that can influence how the context of the message is perceived by others. Speech emotion recognition (SER) is one of a field study that is very intriguing to explore because human-computer interaction (HCI) related technologies such as virtual assistant that are implemented nowadays rarely considered the emotion contained in the information relayed by human speech. One of the most widely used ways to perform SER is by extracting features of speech such as mel frequency cepstral coefficient (MFCC), mel-spectrogram, spectral contrast, tonnetz, and chromagram from the signal and using a one-dimensional (1D) convolutional neural network (CNN) as a classifier. This study shows the impact of implementing a combination of an optimized multi-feature stack and optimized 1D deep CNN model. The result of the model proposed in this study has an accuracy of 90.10% for classifying 8 different emotions performed on the ryerson audio-visual database of emotional speech and song (RAVDESS) dataset.

Corresponding Author:

Muhammad Farhan Fadhil
Department of Computer Science, BINUS Graduate Program, Master of Computer Science
Bina Nusantara University
27, Kebon Jeruk Raya Street, RT.1/RW.9, Kebon Jeruk, Jakarta 11480, Indonesia
Email: muhammad.fadhil009@binus.ac.id

1. INTRODUCTION

Communication is a fundamental element of human life. Communication allows us to receive and deliver messages, exchange knowledge, as well as convey information from one party to another. Over time, the way human communicates evolves in parallel with the advancement of technology. Technology advancement in the digital era is growing exponentially in numbers [1]. In the process of verbal communication, emotion plays a significant role that can influence the way a person receives information verbally as well as determines the receiver's perception [2].

Emotion is a crucial part of communication to clarify the context of an information. However, the current development of technology has not yet implemented emotion recognition in many systems. One of the examples that shows lack of implementation of emotion detection is in virtual assistant that only has the ability to receive raw information from users without taking into account the presence of human emotions in it.

Advancement in speech emotion recognition (SER) might be needed to which allows the detection of human's emotions by computers so that in the future, computers can give the user the appropriate responses to the information given that takes into account the emotions behind it. According to Ekman's theory of emotion categorization, it is stated that the human emotion could be categorized into 6 main emotions, i.e., fear, anger, joy/happy, sadness, disgust, and surprise [3]. However, in this study two additional emotions were added, which are calm and neutral.

What makes this topic intriguing to discuss is because human emotions are subjective. Research shows that the average level of human accuracy in classifying emotions is approximately about 65.8% [4]. For emotion classification in a speech data, we need to first extract the features of the data. Combining several typical features can influence the performance of different classification models [5]–[7]. One method of combining several of these features before they are classified is by doing something called stacking features. From a study conducted in 2022, it is proven that the arrangement of the feature stack can affect the model performance [8]. In this research, the stacking feature orders used are the ones with the best performance that has been researched by Tanoko and Zahra [8], namely spectral contrast, tonnetz, chromagram, mel-spectrogram, and mel frequency cepstral coefficient (MFCC).

The method used in this study will be divided into several sections, which are data fetching, data preprocessing, feature extraction, feature stacking and classification. The dataset used in this study is the ryerson audio-visual database of emotional speech and song (RAVDESS) dataset [9]. The classification method that will be used as the base model is a one-dimensional (1D) convolutional neural network (CNN). This study aims to test the performance of building a SER system by combining an optimized-order multi-feature stack and classification using the 1D CNN Issa *et al.* [7] model which has been fine-tuned. From this study, it is hoped that the resulting of this research can be used as a basis for future studies in this field.

2. RELATED WORK

The establishment of a SER system will be carried out by extracting several features which will then be combined and put into a 1D CNN as a classifier which will produce a prediction of which emotion best describes a particular emotion in a speech data. Studies in the field of SER have achieved many great results in performing emotion classification. This section will describe some of the previous research related to this study.

2.1. The use of different features in speech emotion recognition

The feature selection in the SER will depend on the type of data that will be extracted. Some of the acoustic features that are usually used in SER are time-domain features, frequency-domain features, and statistical features [10]. Time-domain features are frequently used by researchers in this particular field, this is because they are perceived to be the most intuitive features to use. Some examples of time-domain features include short-time zero crossing rate [11], short-time energy [12], and pitch frequency.

Frequency-domain features are the features where in the process of extracting them, the audio signal must first be transformed from time-domain to frequency-domain in order that in the end, the features per frequency can be extracted. This feature is very strongly related to how humans themselves have a perception of emotion in sound, therefore these features are commonly called perceptual features [13]. Some examples of frequency-domain features are MFCC, linear prediction cepstral coefficient (LPCC), perceptual linear predictive cepstrum coefficients (PLPC), and many others. Several previous studies also stated that these perceptual features can be said to be the features that have the best performance in recognizing emotions [14], [15].

Meanwhile, statistical features are the features that are especially used at the utterance level [16]. Different from frequency-domain and time-domain features which are frame level. Utterance level features have the ability to scan the entire emotional speech more deeply [17]. Some examples of statistical features are average, max, min, std deviation, variance, and median.

2.2. Various CNN-based models used in performing speech emotion recognition

One of the CNN-based models used in SER was proposed by Trigeorgis *et al.* [18]. This model is a CNN-based model that uses multi-layer CNNs and finds better performance in terms of accuracy than traditional machine learning methods such as support vector machine (SVM) and random forest [18]. In a 2016 study, Lim *et al.* [19] proposed a better method without using handcrafted features using a combination of CNN and recurrent neural network (RNN) where the model can perform feature learning with long-short term memory (LSTM) and classification using several methods such as CNN and CNN+RNN. It also states that in the speech emotion detection classification process using concatenated CNNs would give more excellent results [19]. Badshah *et al.* [20] conducted a study in 2017 on the Berlin dataset to classify 7 emotions using the spectrogram feature of speech signals that were entered into a deep CNN consisting of 3 CNNs and 3 fully connected layers and get overall accuracy results of 68%.

2.3. Summary and challenges

Based on numerous previous studies in the field of SER itself, various classification methods such as hidden Markov model (HMM) [21], SVM [22], and gaussian mixture model (GMM) [23], deep neural network (DNN)-like LSTM model [24] and CNN [7] show that the CNN architecture is a good option in SER even though the DNN architecture proposed by Mannepalli *et al.* [25] itself has really high accuracy at 99.17%. However, the number of samples and labels used are not specified in the adaptive fractional deep belief network (AFDBN) method that is proposed. Various features used in different studies also have diverse results depending on the feature extraction method used. However, it can also be concluded from the results of various studies that perceptual/frequency-domain features can have a very positive impact on the SER model training process [15].

3. METHOD

This work uses the Python programming language and uses the librosa library for processing the speech signal and feature extraction. Past studies claim the librosa toolkit has better feature set performance than other toolkits such as pyAudioAnalysis and Geneva minimalistic acoustic parameter set (GeMAPS) [26]. Figure 1 shows a diagram of the process flow on how the SER will be carried out in this study. The method used in this research is divided into 7 stages, including: data fetching, data preparation, data preprocessing, framing and windowing, feature extractions, feature stacking, and classification model development.

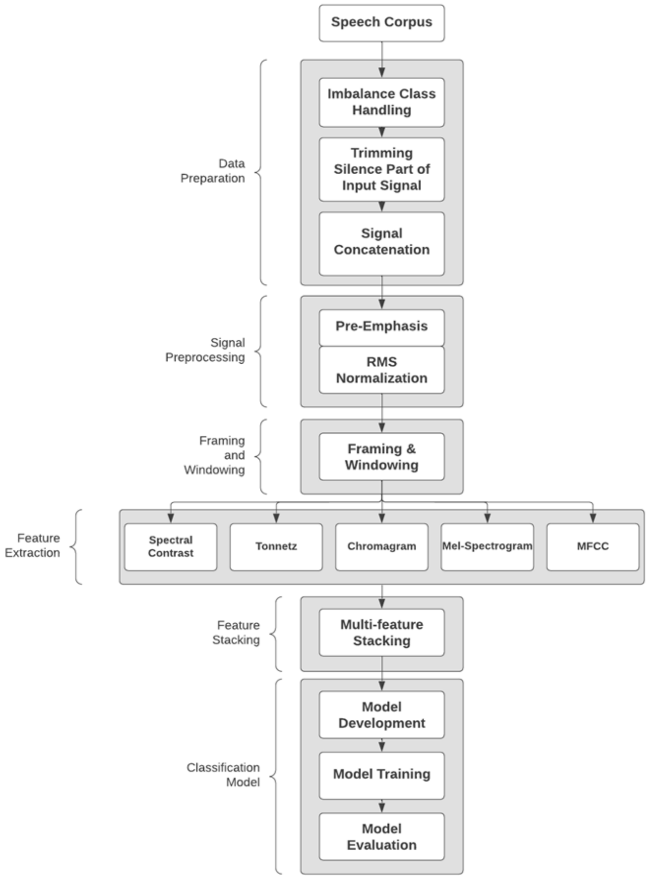

Figure 1. Proposed SER process flow

3.1. Data fetching

The dataset used in this study is the RAVDESS dataset [9]. This emotional speech corpus contains a total of 1440 audio-only data consisting of 24 different actors and 8 classes of emotions consisting of calm, happy, sad, angry, fearful, surprise, and disgust. Figure 2 shows the distribution of data from each class of emotions from the dataset. According to the data distribution shown, it indicates that there is a slight class imbalance that occurs in ‘neutral’ emotions where the class only has 96 data besides the other classes, which each have 192 data.

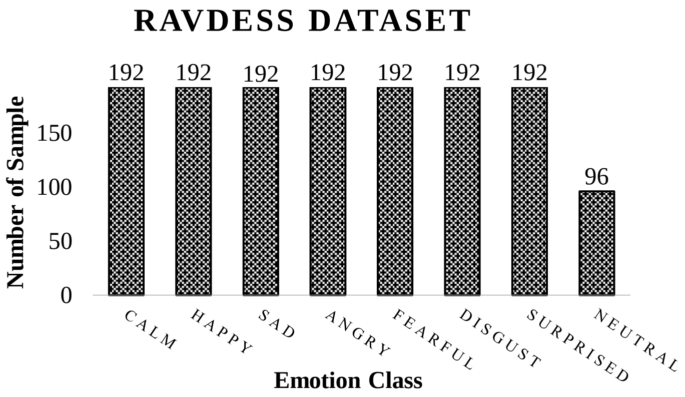

Figure 2. Data count for each class in the dataset

3.2. Data preparation

The data preparation is carried out in 3 main steps, such as: imbalance class handling, trimming input signal, and signal concatenation. The steps of data preparation are done in order to achieve a great SER performance. In SER, performing a proper data preparation in speech signal could also help the model to achieve better result in performing emotion classification.

3.2.1. Imbalance class handling

When handling imbalanced datasets, data resampling can be done using several methods such as undersampling and oversampling [27]. In the case of speech signals, both methods can be used to perform balancing, but the method that will be used in this research is the undersampling method. The undersampling method is done by randomly eliminating the majority data until it reaches the number of minority class data [27]. After the undersampling process, the final total sample obtained is 96 data for each class.

3.2.2. Trimming input signal

Trimming each silent part of the speech signal is done to ensure that the signal that will be processed for feature extraction will be cleaner and expected that it can eliminate random noise, which may have a negative impact on the model performance. This trimming process includes a validation that ensures the signal output from this trimming process produces a signal with a 2 second duration.

3.2.3. Signal concatenation

After all the signals have been trimmed, for each signal that belongs to a specific class, a concatenation process will be carried out, which will combine all data in a class into one long signal. This concatenation was done to ensure that each sample in a signal will not be wasted, which has previously been proven to improve the performance of the SER model [28].

3.3. Data preprocessing

The data preprocessing process is recommended to be carried out, especially for datasets that have different speakers and recording levels [29]. This is done to normalize the features so that during the feature extraction process, all data will have equivalent features [29]. In this study, the two preprocessing methods used were pre-emphasis and root mean square (RMS) normalization.

3.3.1. Pre-emphasis

Pre-emphasis is one of the signal preprocessing methods, where the components of high frequencies are emphasized (boosted) [30]. The goal of pre-emphasis is to improve the signal-to-noise ratio (SNR) of the input signal by reducing the level of low-frequency noise and improving the quality of the signal. Generally, this method is done by applying a high-pass filter which boosts the high frequency of a signal. One of the most frequently used pre-emphasis methods is the first order high-pass filter [30]. Assuming $s(n)$ is the input signal, α with the value of 0.94 as a constant that determines the cutoff frequency of the single-zero filter through which $s(n)$ passes [30]. The first order high-pass filter can be formulated as (1):

$$y(n) = s(n) - \alpha s(n-1) \qquad (1)$$

3.3.2. Root mean square normalization

RMS normalization is a technique in signal processing that is used to adjust the overall level of a signal [31]. The process involves calculating the RMS value of the signal and using this value to scale the entire signal to a target value. This method of normalization makes sure that the input signal has a constant power and amplitude throughout the entire signal. The RMS value is formulated as (2):

$$RMS \triangleq \sqrt{\sum_{n=1}^{N} x^2(n)} \tag{2}$$

where N is the number of samples and x is the input signal:

$$x_{norm} = \frac{x}{RMS} \tag{3}$$

Then the original signal x is divided by *RMS* value to obtain the normalized signal (*x_norm*) [31].

3.4. Framing and windowing

Framing and windowing are common techniques used in speech processing. Framing refers to the process of dividing a continuous signal into smaller segments called frames [32]. Each frame then will be processed individually. In this case, feature will be extracted for each individual frame. On the other side, windowing is a process of applying a weighting function with the aim to reduce the effects of discontinuities at the edge of the frames. Windowing is a mandatory process because the frames we obtain from the framing process doesn't have something to connect each frame. The most commonly used windowing function are the hanning window, hamming window, and blackman window [33].
The equation of the hanning window function is given by:

$$w(n) = 0.5 - 0.5 \times \cos(\frac{2\pi n}{(N-1)}) \tag{4}$$

where $w(n)$ is the value of the hanning window at sample point n, N is the total number of samples in the window and π is the mathematical constant *pi*.

Figure 3 shows the visualization of the framing and windowing process. In this study, the *frame_length* that will be used in the framing process is 2 s/44.100 frames (since the sample rate of the data is 22.050) and the windowing process used is the hanning window [33]. Therefore, a single class obtains a total of 191 slices achieved by 96 slices of the framing process added with 95 slices of the windowing process.

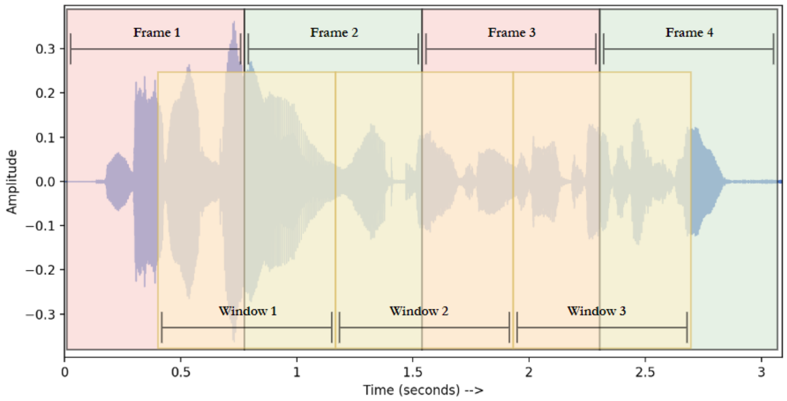

Figure 3. Visualization of framing and windowing process

3.5. Feature extraction

After the framing and windowing process, a total of 5 features will be extracted from each slice. The features that will be taken from each slice are spectral contrast, tonnetz, chromagram, mel-spectrogram, and MFCC since it is widely used for SER [34]. Spectral contrast refers to the difference in intensity between adjacent frequencies or wavelength in a spectrum. It can be measured by calculating the standard deviation of the amplitudes in a frequency band. The spectral contrast plays a significant role in determining the perceived similarity between sounds [35].

Tonnetz are added as a feature extraction method in this research because although mel-spectrogram and MFCC are good for identifying and tracking the timbre fluctuations but tend to have a poor performance in distinguishing pitch classes and harmonies. The tonnetz are used to map out relationships of pitches and harmonies. Chromagram is a form of visual representation of the pitch classes of an audio signal. It is often used in musical information retrieval that represents the harmonic content of a signal [36]. The tonnetz and the chromagram feature both added to compliment each other since they are both are used in the representation of harmony and pitch to balance out the 'timbre-only' representation such as MFCC and mel-spectrogram.

The mel-spectrogram is a type of spectrogram that represents an audio signal in respect of a mel frequency scale which is a non-linear scale that mimics the human auditory system. MFCC is a feature of audio signal that is used to represent the spectral envelope of speech or music signal in a compact form. This feature is by far the most similar to human auditory perception among the other features used in this study. The MFCC is obtained by taking the power spectrum of an audio signal and applying it to a mel-filter bank whose output then is applied to some logarithmic block followed by computation of inverse discrete cosine transform (IDFT) [6].

The hyperparameter for feature extraction used in this study is the following: MFCC uses a of the filter band size of 40. For the mel-spectrogram uses a hop length of 512, windowing method of Hann window and a mel size of 128. For spectral contrast uses a band size of 7, with a hop length of 512, windowing method of Hann window. The chromagram uses hop length of 512, windowing method of Hann window, and chromagram size of 12. The number of features taken from the extraction method is shown in Table 1. The number of coefficients taken are common numbers used in previous studies [8]. These features then will be stacked into a 1D array that will have the length of the sum of *n* which in this case will be 193.

Table 1. Features extracted for each signal slice [8]

Speech feature	Number of coefficient (*n*)
Spectral contrast	7
Tonnetz	6
Chromagram	12
Mel-spectrogram	128
MFCC	40

3.6. Feature stacking

The next step is concatenating all features extracted into a 1D array. Since all the features have different shapes and sizes, by taking the mean value of each time axis, it can then be stacked into a single array. As proven in previous study, the order of which these features are stacked could impact the performance of the model [8]. Therefore, Tanoko and Zahra [8] performed a study and found that the best order of the multi-feature stack in order is spectral contrast, tonnetz, chromagram, mel-spectrogram, followed by MFCC. This study uses the best feature order based on previously mentioned study [8].

3.7. Classification model

The classification model used in this study is a 1D CNN that is initially based on [7] with some tweaking. The initial layer receives the stacked features with the size of 193. The first convolutional layer uses a filter size of 256 and a kernel size of 8 with the stride of 1. Then the result is activated using an activation layer with the type of rectifier linear units (ReLU). Then, it is followed by a dropout layer with the dropout rate of 0.5. The next convolutional layer has the same parameter as the previous convolutional layer but this time, the output is taken to a batch normalization which then activated by the activation of ReLU followed with the dropout rate of 0.5. The third convolutional layer with filter size of 128 with kernel size of 8. The output is then activated using ReLU and taken to the next convolutional layer that has the same parameter as the third convolutional layer but this time the output will be applied to a 1D max-pooling layer with the pool size of 4. The output of this max-pooling layer is then inserted to the next convolutional layer with the filter size of 64 and kernel size of 5 and activated using ReLU. Then, it is followed by a dropout rate of 0.1 which output's is taken to the next 1D max-pooling layer with the pool size of 4 then inserted to the last convolutional layer with filter size of 256 and kernel size of 5 and activated using ReLU followed by a dropout rate of 0.1. The output is then flattened using a flatten layer which is then followed by a dense layer of size 8 representing the number of emotions that will be classified with a softmax activation function. Figure 4 shows a diagram of the classification model used in this study which is a 1D deep CNN model.

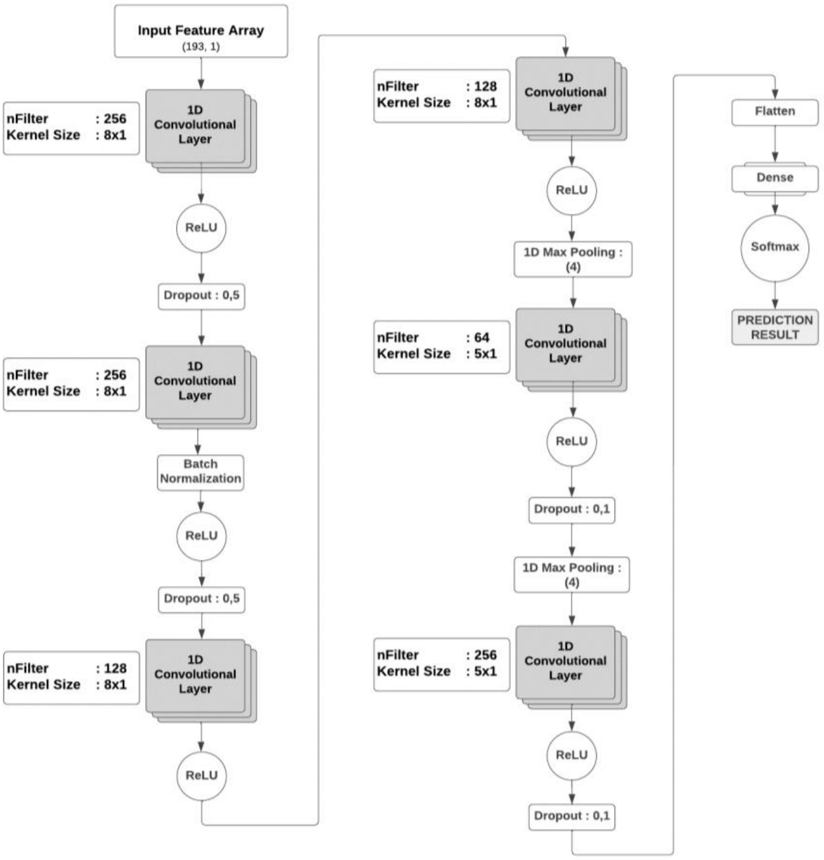

Figure 4. Proposed model of the deep CNN used

As seen in Figure 4, the 1D CNN model is based on the Issa *et al.* [7] deep CNN model. The differences in the model used in this research compared to model used by [7] are the following:

- The kernel size of the first four convolutional layer is (8×1).
- Filter size differences of the 1D convolutional layer such as: second layer *filter_size* is 256; fourth layer *filter_size* is 64; last layer *filter_size* is 256.
- Having batch normalization right after the second convolutional layer.
- Having a dropout layer with the rate of 0.5 in between the first 1D convolutional layer and second 1D convolutional layer.
- Dropout rate after the second convolutional layer is 0.5.
- 1D max pooling with the pool_size of 4 is done after the fifth convolutional layer.
- Added another 1D max pooling layer of size 4 after the sixth convolutional layer.
- Removed the 1D max pooling layer between the second and third convolutional layer.

This model uses an adam optimizer with the learning rate of 0.0001 and epsilon of 1*e*-07. The model also uses a loss function of ‘categorical_crossentropy’ since in this case this model is used for classification of more than 2 classes. Then the model is trained with the batch_size of 32 with 200 epochs. The training process also uses a callback function ReduceLROnPlateau [37] that monitors the validation loss with a factor of 0.8, patience of 15, verbose of 1 and minimum learning rate 1*e*-8. This callback function is used to reduce the learning rate whenever the validation loss has stopped improving.

4. RESULTS AND DISCUSSION

4.1. Results

Table 2 shows the final accuracy, precision, recall, and F1 score of the model used in this study. The performance of the model shows a pretty significant improvement compared to the baseline model from previous work of Issa *et al.* [7] which also uses 1D CNN. This study proves that by using the same 5 features and using the same base model as used in [9] by using an optimized multi-feature stack and fine-tuning, a significant improvement could be made. Figure 5 is a chart representing the accuracy of each emotion class prediction, which shows that the emotions class ‘surprised’, ‘fearful’, and ‘calm’ as the top 3 most accurately predicted emotions, and the ‘sad’ emotion as the least accurately predicted class. The overall accuracy of the model averages 90.1%.

Table 2. Model evaluation result

Evaluation metrics	Score (%)
Accuracy	90.1
Precision	91.3
Recall	90.1
F1	90.9

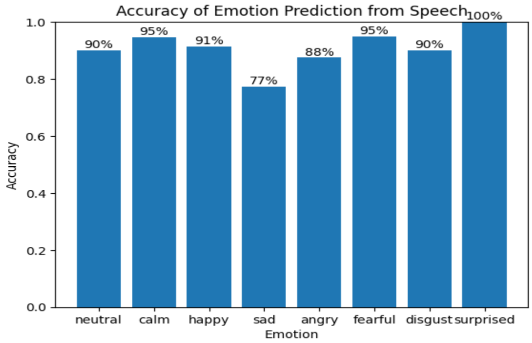

Figure 5. Prediction accuracy for each emotion class

4.2. Discussion

In this work, an optimized deep CNN architecture is proposed. The proposed architecture shows a state-of-the-art result with better performance in terms of accuracy in classifying eight emotions from the RAVDESS dataset, compared to previous work that also uses CNN-based model. This study proves, by combining the optimized steps of preprocessing, framing, windowing, applying an optimized multi-feature order stack and tweaking the deep CNN model, could result in improved performance of previous models.

However, the analysis made in this research relied solely on a single dataset which is the RAVDESS dataset. While the RAVDESS dataset offers diverse emotional speech data, its sole usage could potentially introduce bias only for the mentioned dataset. Another limitation in this study is where this study only focuses on performing classification of 8 distinct emotion classes. Future work may try to incorporate the optimization steps on different datasets with different range emotion classes to determine the effectiveness of this proposed method on different domains. Table 3 is a comparison of different methods used in performing SER that use similar methods and use the same dataset. As shown from Table 3, the proposed model outperforms other models that uses CNN-based model on performing classification of 8 emotions from the RAVDESS dataset, by combining an optimized steps (as mentioned in section 3.6 and 3.7) in the development of the model architecture.

Table 3. Comparison with previous work with similar methods

Previous work	Method	Accuracy (%)
Issa *et al.* (2020) [7]	1D CNN+5 features	71.61
Zeng *et al.* (2019) [38]	1D CNN+spectrogram	65.97
Mustaqeem and Kwon (2020) [39]	2D CNN+spectrogram	79.50
Tanoko and Zahra (2022) [8]	1D CNN+5 features (optimized)	79.17
Sultana *et al.* (2022) [40]	Deep CNN with time-distributed flatten layer and bidirectional long short-term memory layer (DCTFB) + spectrogram	82.70
Proposed method	1D CNN (optimized)+5 features (optimized)	90.10

5. CONCLUSION

In conclusion, SER is a challenging yet intriguing task to tackle. Piecing every single component in the SER architecture is also a challenge for researchers. Finding the right amount of balance of every component in the architecture such as the preprocessing method, the feature extraction, the classification model itself, and many other steps, is mandatory to have a great model and achieving great result. This research uses an optimized multi-feature stack in combination with fine-tuning the classification model which greatly impacts the performance of SER. After performing optimization in each different steps of the SER, the fine-tuned 1D CNN model achieved the highest accuracy of 90.10% in performing classification for 8 different emotion classes.

Future work that could be done in advancing this work is to try work multi-corpus and having them complement each other and combining other classification method to hopefully could further improve the studies in this field. This study also only uses a single classification method as a baseline model which is the CNN model. Further research could also incorporate the use of different models such as the HMM, GMM, or the RNN model.

ACKNOWLEDGEMENT

The research presented in this paper, titled "speech emotion recognition with optimized multi-feature stack using deep convolutional neural networks" was made possible through the financial support provided by Bina Nusantara University.

REFERENCES

[1] A. Kumar and N. Epley, "It's Surprisingly Nice to Hear You: Misunderstanding the Impact of Communication Media Can Lead to Suboptimal Choices of How to Connect With Others," *Journal of Experimental Psychology: General*, vol. 150, no. 3, pp. 595–607, Mar. 2021, doi: 10.1037/xge0000962.

[2] M. Tsvetkova, "The Non-Linguistic Context–a Bridge To Linguistic Items and Phenomena," *Studies in Linguistics, Culture, and FLT*, vol. 02, pp. 219–226, 2017, doi: 10.46687/silc.2017.v02.018.

[3] P. Ekman, "An Argument for Basic Emotions," *Cognition and Emotion*, vol. 6, no. 3–4, pp. 169–200, May 1992, doi: 10.1080/02699939208411068.

[4] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, Nov. 2003, doi: 10.1016/S0167-6393(03)00099-2.

[5] Y. Zhou, Y. Sun, J. Zhang, and Y. Yan, "Speech emotion recognition using both spectral and prosodic features," in *Proceedings-2009 International Conference on Information Engineering and Computer Science,* Dec. 2009, pp. 1–4, doi: 10.1109/ICIECS.2009.5362730.

[6] S. Lalitha, S. Tripathi, and D. Gupta, "Enhanced speech emotion detection using deep neural networks," *International Journal of Speech Technology*, vol. 22, no. 3, pp. 497–510, Sep. 2019, doi: 10.1007/s10772-018-09572-8.

[7] D. Issa, M. F. 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.

[8] Y. Tanoko and A. Zahra, "Multi-feature stacking order impact on speech emotion recognition performance," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 6, pp. 3272–3278, Dec. 2022, doi: 10.11591/eei.v11i6.4287.

[9] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north American english," *PLoS One*, vol. 13, no. 5, p. e0196391, May 2018, doi: 10.1371/journal.pone.0196391.

[10] C. Wang, Y. Ren, N. Zhang, F. Cui, and S Luo, "Speech Emotion Recognition Based on Multi-feature and Multi-lingual Fusion," *Multimed Tools Applications*, vol. 81, no. 4, pp. 4897–4907, 2022, doi: 10.1007/s11042-021-10553-4.

[11] D. Nath and S. K. Kalita, "An effective age detection method based on short time energy and zero crossing rate," in *2014 2nd International Conference on Business and Information Management,* Jan. 2014, pp. 99–103, doi: 10.1109/ICBIM.2014.6970942.

[12] M. Jalil, F. A. Butt, and A. Malik, "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals," in *2013 The International Conference on Technological Advances in Electrical, Electronics and Computer Engineering,* May 2013, pp. 208–212, doi: 10.1109/TAEECE.2013.6557272.

[13] M. Athineos and D. P. W. Ellis, "Frequency-domain linear prediction for temporal features," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding,* 2003, pp. 261–266, doi: 10.1109/ASRU.2003.1318451.

[14] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using Fourier parameters," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 69–75, Jan. 2015, doi: 10.1109/TAFFC.2015.2392101.

[15] X. Mao, L. Chen, and L. Fu, "Multi-level speech emotion recognition based on HMM and ANN," in *2009 WRI World Congress on Computer Science and Information Engineering,* 2009, vol. 7, pp. 225–229, doi: 10.1109/CSIE.2009.113.

[16] V. A. Petrushin, "Emotion recognition in speech signal: Experimental study, development, and application," in *6th International Conference on Spoken Language Processing,* Oct. 2000, vol. 2, pp. 222-225-0, doi: 10.21437/icslp.2000-791.

[17] S. Furui, "Comparison of Speaker Recognition Methods Using Statistical Features and Dynamic Features," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 342–350, Jun. 1981, doi: 10.1109/TASSP.1981.1163605.

[18] G. Trigeorgis *et al.*, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings,* Mar. 2016, vol. 2016, pp. 5200–5204, doi: 10.1109/ICASSP.2016.7472669.

[19] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2017, pp. 1–4, doi: 10.1109/APSIPA.2016.7820699.

[20] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," in *2017 International Conference on Platform Technology and Service,* Feb. 2017, pp. 1–5, doi: 10.1109/PlatCon.2017.7883728.

[21] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proceedings-IEEE International Conference on Multimedia and Expo*, 2003, vol. 1, pp. I401–I404, doi: 10.1109/ICME.2003.1220939.

[22] 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, pp.

163–168, 2021, doi: 10.14569/IJACSA.2021.0120422.

[23] H. Tang, S. M. Chu, M. Hasegawa-Johnson, and T. S. Huang, “Emotion recognition from speech via boosted Gaussian mixture models,” in *Proceedings-2009 IEEE International Conference on Multimedia and Expo*, Jun. 2009, pp. 294–297, doi: 10.1109/ICME.2009.5202493.

[24] J. Zhao, X. Mao, and L. Chen, “Speech emotion recognition using deep 1D & 2D CNN LSTM networks,” *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019, doi: 10.1016/j.bspc.2018.08.035.

[25] K. Mannepalli, P. N. Sastry, and M. Suman, “A novel Adaptive Fractional Deep Belief Networks for speaker emotion recognition,” *Alexandria Engineering Journal*, vol. 56, no. 4, pp. 485–497, Dec. 2017, doi: 10.1016/j.aej.2016.09.002.

[26] B. T. Atmaja and M. Akagi, "On The Differences Between Song and Speech Emotion Recognition: Effect of Feature Sets, Feature Types, and Classifiers," *2020 IEEE REGION 10 CONFERENCE (TENCON)*, Osaka, Japan, 2020, pp. 968-972, doi: 10.1109/TENCON50793.2020.9293852.

[27] V. S. Spelmen and R. Porkodi, “A Review on Handling Imbalanced Data,” in *Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies*, Mar. 2018, pp. 1–11, doi: 10.1109/ICCTCT.2018.8551020.

[28] F. J. Charpentier and M. G. Stella, “Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation.,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*, 1986, pp. 2015–2018, doi: 10.1109/icassp.1986.1168657.

[29] 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, 2020, doi: 10.1016/j.specom.2019.12.001.

[30] R. Vergin and D. O’Shaughnessy, “Pre-emphasis and speech recognition,” in *Canadian Conference on Electrical and Computer Engineering*, 1995, vol. 2, pp. 1062–1065, doi: 10.1109/ccece.1995.526613.

[31] C. Panagiotakis and G. Tziritas, “A speech/music discriminator based on RMS and zero-crossings,” *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 155–166, Feb. 2005, doi: 10.1109/TMM.2004.840604.

[32] R. W. Schafer and L. R. Rabiner, “Digital Representations of Speech Signals,” *Proceedings of the IEEE*, vol. 63, no. 4, pp. 662–677, 1975, doi: 10.1109/PROC.1975.9799.

[33] F. J. Harris, “On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform,” *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978, doi: 10.1109/PROC.1978.10837.

[34] S. K. Pandey, H. S. Shekhawat, and S. R. M. Prasanna, “Deep learning techniques for speech emotion recognition: A review,” in *2019 29th International Conference Radioelektronika*, Apr. 2019, pp. 1–6, doi: 10.1109/RADIOELEK.2019.8733432.

[35] J. Yang, F. L. Luo, and A. Nehorai, “Spectral contrast enhancement: Algorithms and comparisons,” *Speech Communication*, vol. 39, no. 1–2, pp. 33–46, Jan. 2003, doi: 10.1016/S0167-6393(02)00057-2.

[36] L. Shi, C. Li, and L. Tian, “Music Genre Classification Based on Chroma Features and Deep Learning,” in *10th International Conference on Intelligent Control and Information Processing*, Dec. 2019, pp. 81–86, doi: 10.1109/ICICIP47338.2019.9012215.

[37] A. Al-Kababji, F. Bensaali, and S. Dakua “Scheduling Techniques for Liver Segmentation: ReduceLRonPlateau Vs OneCycleLR,” *Intelligent Systems and Pattern Recognition: Second International Conference, Hammamet, Tunisia, March 24–26, 2022, Revised Selected Papers*, pp. 204–212, 2022.

[38] Y. Zeng, H. Mao, D. Peng, and Z. Yi, “Spectrogram based multi-task audio classification,” *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3705–3722, Feb. 2019, doi: 10.1007/s11042-017-5539-3.

[39] Mustaqeem and S. Kwon, “A CNN-assisted enhanced audio signal processing for speech emotion recognition,” *Sensors (Switzerland)*, vol. 20, no. 1, p. 183, Dec. 2020, doi: 10.3390/s20010183.

[40] S. Sultana, M. Z. Iqbal, M. R. Selim, M. M. Rashid, and M. S. Rahman, “Bangla Speech Emotion Recognition and Cross-Lingual Study Using Deep CNN and BLSTM Networks,” *IEEE Access*, vol. 10, pp. 564–578, 2022, doi: 10.1109/ACCESS.2021.3136251.

BIOGRAPHIES OF AUTHORS

Muhammad Farhan Fadhil is a graduate student currently studying in Bina Nusantara University under the Department of Computer Science majoring in Computer Science. His research interest includes speech technology such as speech emotion recognition, and signal processing. He can be contacted at email: muhammad.fadhil009@binus.ac.id.

Amalia Zahra is a lecturer at the Master of Computer Science, Bina Nusantara University, Indonesia. She received her Bachelor degree in Computer Science from the Faculty of Computer Science, University of Indonesia (UI) in 2008. She does not have a master degree. Her Ph.D. was obtained from the School of Computer Science and Informatics, University College Dublin (UCD), Ireland in 2014. Her research interests cover various fields in speech technology, such as speech recognition, spoken language identification, speaker verification, and speech emotion recognition. Additionally, she also has interest in natural language processing (NLP), computational linguistics, and machine learning. She can be contacted at email: amalia.zahra@binus.edu.