

Multi-feature stacking order impact on speech emotion recognition performance

Yoga Tanoko, Amalia Zahra

Department Computer Science, BINUS Graduate Program, Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Jun 21, 2022

Revised Jul 26, 2022

Accepted Aug 25, 2022

Keywords:

Chromagram

CNN

Mel-spectrogram

MFCC

Spectral contrast

Speech emotion recognition

Tonnetz

ABSTRACT

One of the biggest challenges in implementing SER is to produce a model that performs well and is lightweight. One of the ways is using one-dimensional convolutional neural network (1D CNN) and combining some handcrafted features. 1D CNN is mostly used for time series data. In time series data, the order of information plays an important role. In this case, the order of stacked features also plays an important role. In this work, the impact of changing the order is analyzed. This work proposes to brute force all possible combinations of feature orders from five features: Mel-frequency cepstral coefficient (MFCC), Mel-spectrogram, chromagram, spectral contrast, and tonnetz, then uses 1D CNN as the model architecture and benchmarking the model's performance on the Ryerson audio-visual database of emotional speech and song (RAVDESS) dataset. The results show that changing the order of features can impact overall classification accuracy, specific emotion accuracy, and model size. The best model has an accuracy of 79.17% for classifying 8 emotion classes with the following order: spectral contrast, tonnetz, chromagram, Mel-spectrogram, and MFCC. Finding a suitable order can increase the accuracy up to 16.05% and reduce the model size up to 96%.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Yoga Tanoko

Department Computer Science, BINUS Graduate Program, Master of Computer Science

Bina Nusantara University

Jl. Raya Kebon Jeruk No.27, RT.1/RW.9, Kebon Jeruk, West Jakarta 11480, Indonesia

Email: yoga.tanoko@binus.ac.id

1. INTRODUCTION

Speech emotion recognition (SER) is a field of science that studies how to recognize emotions from speech input. This field is interesting because emotion is subjective. From an utterance, not everyone can correctly identify the type of emotion that is present when the speaker speaks. Research shows that the accuracy of classification performed by humans is 65.8% [1]. By identifying the proper emotions, the response given can be better and the experience of interaction can be improved. The computer is now able to receive and reply to voice commands once the device needed for the task has been installed. SER could be used to enable computers to detect emotions and improve experience in human-computer interaction [2]. Some SER applications have been developed in human-computer interaction, such as robots [3], online learning [4], and psychological consultation [5]. Although it has many applications, SER is still a challenging task because there is no certain way to extract and categorize emotions from speech. In speech data, several features can be retrieved, such as prosodic features, spectral features, audio quality, and Teager energy operator (TEO). The difference in the features used in the SER will affect the quality of a model [6–9].

Several methods have been tested in SER, including classical classification methods such as hidden Markov model (HMM) [10], support vector machine (SVM) [11], and Gaussian mixture model (GMM) [12],

as well as using deep learning such as long-short term memory (LSTM) [13] and convolutional neural network (CNN) [6]. The deep learning approach detects the high-level salient features to achieve better accuracy compared to classical classification methods. From the tests that have been carried out before [8], [14], [15], it is known that the CNN architecture performs better than other methods. The usage of deep CNN boosts the computational complexity of the whole model.

Kwon [15] proposed to modify the stride in 2D convolution layer to mimic pooling layer then remove pooling layer from the model. Spectrogram is extracted from raw speech to be used as input. This results in 79.5% accuracy on the Ryerson audio-visual database of emotional speech and song (RAVDESS) dataset with a model size of 34.5 MB. On the other hand, Issa *et al* [6] proposed combining 1D CNN and low-level handcrafted features to reduce the complexity of the model and reduce unnecessary information from raw speech. They used Mel-frequency cepstral coefficients (MFCCs), Mel-spectrogram, chromagram, spectral contrast, and Tonnetz as features, by taking the mean value along the time axis and then stacking them on each other as input. The result shows 71.61% accuracy. The accuracy can be improved by finding the suitable feature order. The issue is kernel in 1D CNN slides along one dimension, so information order is important, which is why 1D CNN is most likely to be used for time series data [16]. By using multiple features, the order of stacking features plays an important role in the model's performance. The challenges now are to find the best order of stacking features and investigate the impact of different feature orders in 1D CNN.

Due to the issues and challenges, the performance of 1D CNN for SER can be improved by finding the best feature stacking order. This work proposes to brute force on all combinations of the five-feature stack, which are MFCC, chromagram, Mel-spectrogram, spectral contrast, and Tonnetz to find the best order and impact of multi-feature order. The experiments were conducted on standards benchmarked with RAVDESS [17]. Brute force was selected since all possible feature order would be tested and finding the impact of feature order would be easier to perform.

2. METHOD

This study aims to find the impact of multi-feature order by brute force all possible combinations. The method consists of four stages: the dataset preparation including data augmentation to increase the number of samples, feature extraction including parameter configuration for each feature, feature stacking, and models developed from the base model that was used in the previous work [6]. This work uses Librosa as a toolkit since it has better feature set performance in comparison to other tools like GeMAPS and pyAudioAnalysis [18]. Google Colab is used as a platform to run the experiment.

2.1. Dataset preparation

The dataset that will be used in this study is RAVDESS [17], containing audio and visual recordings of 12 men and 12 women who say English sentences with eight different emotional expressions: sad, happy, angry, calm, fearful, surprised, neutral, and disgusted, with a total recording of 1440 samples. In this work, the dataset is split into three partitions, 70% for training, 15% for validation, and 15% for test. In the RAVDESS dataset, the sample distribution for each emotion is fairly even, but for neutral emotions, the total duration is only about 50% of the other emotions (see Figure 1). To overcome the imbalance in the dataset, several methods will be applied to make the dataset more balanced. This can be done by augmenting the data on neutral emotions or by reducing the size of other emotions to balance them with neutral emotions. In this study, data augmentation was applied [19] to multiply the data that will be used to create the model. This issue to the fact that CNN model requires a lot of data to make it more stable. For neutral emotion, random noise was added for augmentation, and then the number of data was balanced. Furthermore, to increase the number of samples, augmentation is performed one more time by combining two processes which are stretching the data with a 0.8 rate and increasing pitch by a factor of 0.7.

2.2. Feature extraction

Features that are used in this paper are MFCC, chromagram, Mel-spectrogram, spectral contrast, and Tonnetz. MFCC and Mel-scaled spectrograms are widely used in SER [20], [21]. These features mimic a certain degree of acceptance of the intrinsic human sound frequency pattern. The MFCC creates a Mel-frequency spectrum, which can be defined as a representation of the short-term sound power spectrum. Fourier transforms and energy spectra were collected and mapped onto a mel-frequency scale. Both Mel-spectrogram and MFCCs are decent in the identification and tracking of timbre fluctuations in a sound file, they tend to be poor in a distinguishable representation of pitch classes and harmony [6]. To handle such situation, chromagram and Tonnetz are added. Chromagram and Tonnetz are similar with respect to the representation of harmony and pitch classes [22], [23]. Spectral Contrast represents the energy contrast computed by comparing the peak energy and valley energy in each band converted from spectrogram frames

[24]. In this study, hyperparameters on the MFCC are used with a filter band size of 40. For the Mel-spectrogram, a mel size of 128 is used, a hop length of 512 with a windowing method in the form of a Hann window. For spectral contrast, a band size of 6 is used, with a hop length of 512, with a windowing method in the form of a Hann window. For chroma, it uses chromagram size of 12 hops length of 512 with a windowing method in the form of a Hann window. As for the Tonnetz feature, the input is a chromagram feature that has been calculated previously.

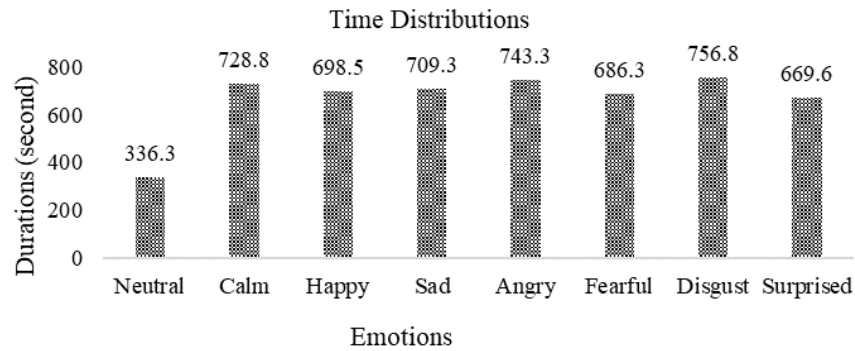


Figure 1. Time distribution of each emotion

2.3. Feature stacking

After feature extraction, the next step is to concatenate all features into one array. Since every feature has a different size, all features are compressed into a one-dimensional array by taking the mean value along the time axis and then stacked on each other. An illustration of the stacking process can be seen in Figure 2 (a) to (c). This process will be repeated for all feature combinations.

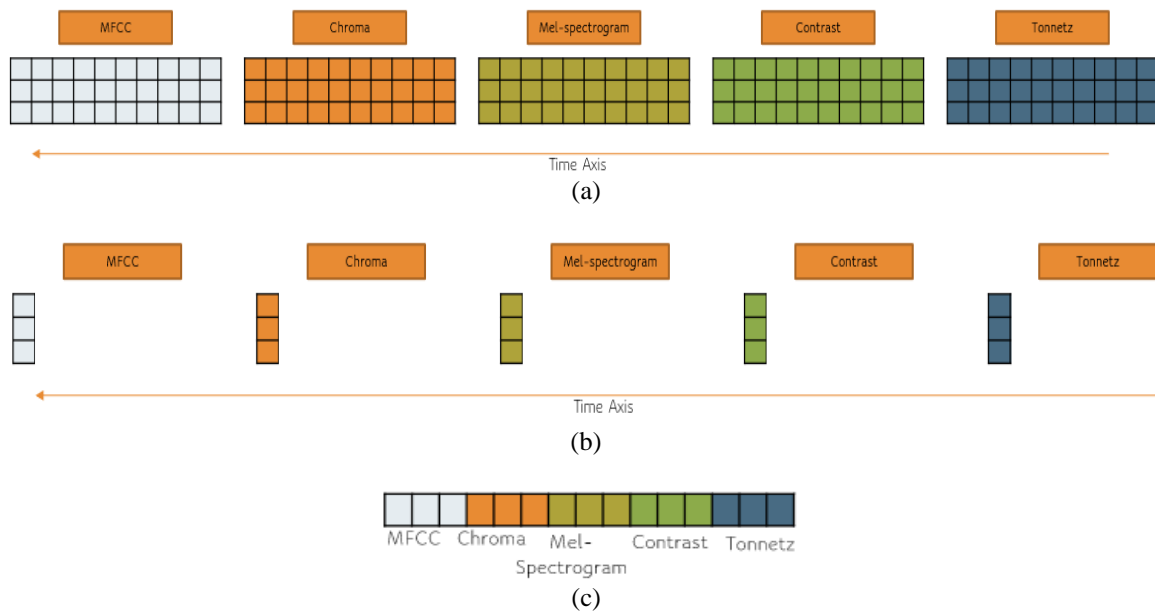


Figure 2. Stacking process of (a) extracted features in a two-dimensional array, (b) compressed into one-dimensional array by taking mean a long-time axis then, and (c) concatenate into one array

2.4. Model development

The model in this study is built using the 1D CNN architecture using a baseline from Issa *et al.* [6] with some adjustments. The first layer receives 193 number arrays as input data. The initial convolution layer

is composed of 256 filters with a kernel size of 8 and a stride of 1. Its output is activated by the rectifier linear units (ReLU) layer. The next convolutional layer has the same number of parameters but is followed by batch normalization before being activated with ReLU. After that, dropout with a rate of 0.25 is applied. The next 3 convolution layers with 128 filters of size 8 are located. Each convolutional layer is followed by ReLU activation layers. The next layer is a convolutional layer with a 128 filter of size 8, followed by batch normalization before being activated with ReLU. The output of this layer is followed by dropout with a rate of 0.25. Then a max-pooling layer with a pool size of 8 is applied. The next two convolution layers with 64 filters of size 8 are located. The output of each convolutional layer is activated with ReLU. The output of this layer is followed by a flattening layer. The output of the flattening layer is received by a fully connected layer with eight units representing eight classes of emotions and activated with softmax activation function. This model uses the RMSProp optimizer with a learning rate of 0.00001 and a decay rate of $1e-6$.

The results of the stacking of features are then entered into the model to be trained. To make sure all the model gets the same treatment, all the training, validation, and test data that are used are the same for every model. For the training process, a batch size of 64 with 300 epochs is used and callback ReduceLROnPlateau [25] is used to monitor loss, and then the learning rate is adjusted by the factor of 0.8, patience of 15, and then a minimum learning rate of 0.000001 is used to make sure the learning rate does not go below 0.000001. Previous research [26] showed that using a learning rate of 0.000001 on CNN architectures to obtain a model with the lowest loss.

3. RESULTS AND DISCUSSION

3.1. Result

The accuracy distribution of all models can be seen in Figure 3. The order of features has a big impact on model performance. From the data collected, the difference between the highest and the lowest accuracy is 16.05%. From Table 1, the highest accuracy is obtained with the following order of features: spectral contrast, Tonnetz, chromagram, Mel-spectrogram, and MFCC. From the top five models, it can be seen that five of them have similarities where Mel-spectrogram and MFCC are placed side by side, while sometimes contrast and tonnetz are also placed side by side. On the other hand, from Table 2, the bottom five models have MFCC and Mel-spectrogram placed far apart. There is no big impact of spectral contrast position on classification performance.

Feature order affects not only overall accuracy, but also accuracy on specific emotions. From Table 3, the result shows that putting specific features close to each other can determine the performance of recognizing specific emotions. Angry and calm emotions have differences in the order of chromagram and Mel-spectrogram. Putting chromagram next to Tonnetz instead of Mel-spectrogram has better performance to recognize angry. On the other side, putting Mel-spectrogram after Tonnetz can result in better accuracy in detecting calm emotion. Fearful and happy emotions have differences in the order of spectral contrast and chromagram. Placing spectral contrast after MFCC yields better performance on recognizing happy while putting chromagram next to MFCC yields better performance on detecting fearful emotion. For happy and sad emotions, the difference is in the order of MFCC and spectral contrast, while putting spectral contrast after Mel-spectrogram has better accuracy at detecting sad. Not only affecting the model's performance, the feature order also affects the model size. The size range of models after training is between 10 MB to 293 MB, while the best model with the order MFCC, Tonnetz, spectral contrast, chromagram, and Mel-spectrogram has a size of 29.6 MB.

Table 1. The order of five top accuracies

Feature order	Accuracy (%)
Spectral contrast, Tonnetz, chromagram, Mel-spectrogram, MFCC	79.17
Tonnetz, chromagram, spectral contrast, Mel-spectrogram, MFCC	78.86
Spectral contrast, Mel-spectrogram, MFCC, chromagram, Tonnetz	78.70
Spectral contrast, chromagram, Mel-spectrogram, MFCC, Tonnetz	78.40
Tonnetz, chromagram, Mel-spectrogram, MFCC, spectral contrast	77.93

Table 2. The order of bottom five accuracies

Feature order	Accuracy (%)
MFCC, Tonnetz, spectral contrast, chromagram, Mel-spectrogram	66.51
Mel-spectrogram, chromagram, MFCC, Tonnetz, spectral contrast	66.51
Chromagram, MFCC, spectral contrast, Tonnetz, Mel-spectrogram	66.20
MFCC, spectral contrast, chromagram, Mel-spectrogram, Tonnetz	65.43
MFCC, spectral contrast, Tonnetz, Mel-spectrogram, chromagram	63.73

Table 3. Best feature order for a specific emotion

Emotion	Feature order	Accuracy (%)
Angry	Spectral contrast, MFCC, tonnetz, chromagram, Mel-spectrogram	86.60
Calm	Spectral contrast, MFCC, tonnetz, Mel-spectrogram, chromagram	98.80
Disgust	Mel-spectrogram, chromagram, spectral contrast, Tonnetz, MFCC	85.39
Fearful	Tonnetz, Mel-spectrogram, MFCC, chromagram, spectral contrast	88.89
Happy	Tonnetz, Mel-spectrogram, MFCC, spectral contrast, chromagram	87.34
Neutral	Spectral contrast, Mel-spectrogram, Tonnetz, chromagram, MFCC	84.78
Sad	Tonnetz, Mel-spectrogram, spectral contrast, MFCC, chromagram	91.95
Surprised	Tonnetz, chromagram, Mel-spectrogram, spectral contrast, MFCC	97.67

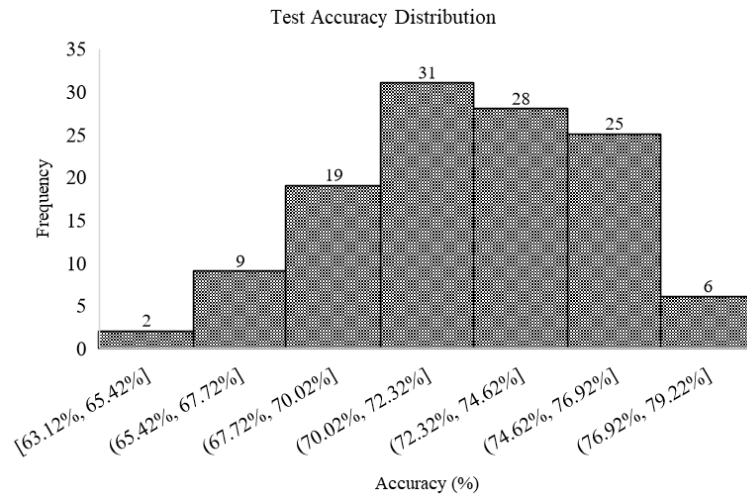


Figure 3. Test accuracy distribution of 120 model on the RAVDESS dataset

3.2. Discussion

In this work, optimized feature order was found by brute force all possible combinations of feature stacking order. 1D CNN is more compact than 2D CNN that is usually used for SER. The main issue of 1D CNN is the order of feature affects model performance. Finding the best stacking order can improve 1D CNN performance to match up with 2D CNN while reducing complexity of the model. Comparing to previous work [6], this work put more focus on finding the best order while the previous work did not. Focusing more on finding the best stacking order shows big improvement in model performance.

The highest accuracy is 79.17%, which is obtained from the feature stacking order of spectral contrast, Tonnetz, chromagram, Mel-spectrogram, and MFCC. Table 4 shows comparison between the result of the work presented in this paper with that from the previous work [6], where the best accuracy is 71.61% with the following order: MFCC, chromagram, Mel-spectrogram, contrast, and Tonnetz. The difference in best accuracy happened because the previous work did not properly find the best order. While in this work, all possible combinations of feature order are experimentally tried to find the best feature order.

Table 4. Comparison with previous work

Previous work	Method	Accuracy (%)	Model size
Kwon [15]	2D CNN+spectrogram	79.5	34.5 MB
Issa <i>et al.</i> [6]	1D CNN+feature order MFCC, chromagram, Mel-spectrogram, spectral contrast, Tonnetz	71.61	-
This work	1D CNN+optimized feature order spectral contrast, Tonnetz, chromagram, Mel-spectrogram, MFCC	79.17	29.6 MB

Beside impact on accuracy, feature order also has an impact on model size. In this work, a smaller model than that from the previous work [15] has been successfully achieved. Our best model has an accuracy of 79.17% and a size of 29.6 MB while previous work results in 79.5% accuracy on the RAVDESS dataset with a model size of 34.5 MB. It shows that the model obtained in this study achieves slightly lower accuracy, but a smaller model size. This happened since 1D CNN was used where it was simpler than 2D CNN.

4. CONCLUSION

Multi-feature usage on SER is a challenging task because many features can be extracted from the speech signal. Each feature has similarities and differences in extracted information. Changing the order of features can have an impact on classification performance, especially using the 1D CNN architecture. This work intends to find the impact of multi-feature order on SER performance on 1D CNN. From the result, it can be concluded that the order of features affects not only the overall accuracy of the model, but also the performance in recognizing specific emotions and the model size. This work achieves better feature order for the RAVDESS dataset with 79.17% accuracy and produces a model with a smaller size of 29.6 MB. Future work can be conducted by using different features like Teager energy to increase the number of features.

ACKNOWLEDGEMENT

This work is supported by the Directorate General of Strengthening for Research and Development, Ministry of Research, Technology, and Higher Education, Republic of Indonesia, as a part of penelitian tesis magister (PTM) research grant to Binus University entitled “Optimasi multi fitur pada sistem pengenalan emosi suara” or “Multi-feature optimization on speech emotion recognition” with contract number: 410/LL3/AK.04/2022, 17th June 2022.




REFERENCES

- [1] 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.
- [2] M. E. Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, Mar. 2011, doi: 10.1016/j.patcog.2010.09.020.
- [3] X. Huahu, G. Jue, and Y. Jian, “Application of Speech Emotion Recognition in Intelligent Household Robot,” in *2010 International Conference on Artificial Intelligence and Computational Intelligence*, Oct. 2010, pp. 537–541. doi: 10.1109/AICI.2010.118.
- [4] L. Cen, F. Wu, Z. L. Yu, and F. Hu, “A Real-Time Speech Emotion Recognition System and its Application in Online Learning,” in *Emotions, Technology, Design, and Learning*, Elsevier, 2016, pp. 27–46. doi: 10.1016/B978-0-12-801856-9.00002-5.
- [5] H. -C. Li, T. Pan, M. -H. Lee, and H. -W. Chiu, “Make Patient Consultation Warmer: A Clinical Application for Speech Emotion Recognition,” *Applied Sciences*, vol. 11, no. 11, p. 4782, May 2021, doi: 10.3390/app11114782.
- [6] D. Issa, M. F. Demirci, and A. Yazici, “Speech emotion recognition with deep convolutional neural networks,” *Biomedical Signal Processing and Control*, vol. 59, May 2020, doi: 10.1016/j.bspc.2020.101894.
- [7] P. Shen, Z. Changjun, and X. Chen, “Automatic Speech Emotion Recognition using Support Vector Machine,” in *Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology*, Aug. 2011, pp. 621–625. doi: 10.1109/EMEIT.2011.6023178.
- [8] H. Patni, A. Jagtap, V. Bhojar, and A. Gupta, “Speech Emotion Recognition using MFCC, GFCC, Chromagram and RMSE features,” in *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, Aug. 2021, pp. 892–897. doi: 10.1109/SPIN52536.2021.9566046.
- [9] Y. Li, C. Baidoo, T. Cai, and G. A. Kusi, “Speech Emotion Recognition Using 1D CNN with No Attention,” in *2019 23rd International Computer Science and Engineering Conference (ICSEC)*, Oct. 2019, pp. 351–356. doi: 10.1109/ICSEC47112.2019.8974716.
- [10] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov model-based speech emotion recognition,” in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, 2003, vol. 2, pp. II-1–4. doi: 10.1109/ICASSP.2003.1202279.
- [11] 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.
- [12] P. Patel, A. Chaudhari, R. Kale, and M. A. Pund, “Emotion Recognition From Speech With Gaussian Mixture Models & Via Boosted Gmm,” *International Journal of Research In Science & Engineering*, 2017.
- [13] 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, Jan. 2019, doi: 10.1016/j.bspc.2018.08.035.
- [14] 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 (PlatCon)*, Feb. 2017, pp. 1–5. doi: 10.1109/PlatCon.2017.7883728.
- [15] Mustaqeem and S. Kwon, “A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition,” *Sensors*, vol. 20, no. 1, p. 183, Dec. 2019, doi: 10.3390/s20010183.
- [16] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, “1D convolutional neural networks and applications: A survey,” *Mechanical Systems and Signal Processing*, vol. 151, p. 107398, Apr. 2021, doi: 10.1016/j.ymsp.2020.107398.
- [17] 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.
- [18] B. T. Atmaja and M. Akagi, “On the Differences Between Song and Speech Emotion Recognition: Effect of Feature Sets, Feature Types, and Classifiers,” in *2020 IEEE REGION 10 CONFERENCE (TENCON)*, Nov. 2020, pp. 968–972. doi: 10.1109/TENCON50793.2020.9293852.
- [19] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, “CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation,” Feb. 2018, doi: 10.21437/SMM.2018-5.
- [20] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, “Speech based human emotion recognition using MFCC,” in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Mar. 2017, pp. 2257–2260. doi: 10.1109/WiSPNET.2017.8300161.
- [21] 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 (RADIOELEKTRONIKA)*, Apr. 2019, pp. 1–6. doi: 10.1109/RADIOELEK.2019.8733432.
- [22] C. Harte, M. Sandler, and M. Gasser, “Detecting harmonic change in musical audio,” in *Proceedings of the 1st ACM workshop on Audio and music computing multimedia - AMCMM '06*, 2006, p. 21. doi: 10.1145/1178723.1178727.
- [23] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai, “Music type classification by spectral contrast feature,” in *Proceedings. IEEE International Conference on Multimedia and Expo*, 2002, pp. 113–116. doi: 10.1109/ICME.2002.1035731.
- [24] Z. Dair, R. Donovan, and R. O’Reilly, “Linguistic and Gender Variation in Speech Emotion Recognition using Spectral Features,” Dec. 2021.
- [25] A. Al-Kababji, F. Bensaali, and S. P. Dakua, “Scheduling Techniques for Liver Segmentation: ReduceLRonPlateau Vs OneCycleLR,” Feb. 2022.
- [26] R. Ren, S. Zhang, H. Sun, and T. Gao, “Research on Pepper External Quality Detection Based on Transfer Learning Integrated with Convolutional Neural Network,” *Sensors*, vol. 21, no. 16, p. 5305, Aug. 2021, doi: 10.3390/s21165305

BIOGRAPHIES OF AUTHORS



Yoga Tanoko    is currently a graduate student in Computer Science Department of Bina Nusantara University. He received his bachelor’s degree in Information Technology from Dinamika Bangsa University in 2019. His research interest includes speech technology such as Speech Emotion Recognition. He can be contacted at email: yoga.tanoko@binus.ac.id



Amalia Zahra    is a lecturer at the Master of Information Technology, Bina Nusantara University, Indonesia. She received her bachelor’s degree in computer science from the Faculty of Computer Science, University of Indonesia (UI) in 2008. She does not have a master’s degree. Her PhD 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, speech emotion recognition, and so on. Additionally, she also has interest in natural language processing (NLP), computational linguistics, machine learning, and artificial intelligence. She can be contacted at email: amalia.zahra@binus.ac.id.