Bulletin of Electrical Engineering and Informatics
Vol. 13, No. 3, June 2024, pp. 1920~1926
ISSN: 2302-9285, DOI: 10.11591/eei.v13i3.6049

Enhancing speech emotion recognition with deep learning using multi-feature stacking and data augmentation

Khasyi Al Mukarram, M. Anang Mukhlas, 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 Jul 25, 2023
Accepted Nov 14, 2023

Keywords:

Convolutional neural network
Data augmentation
Multi-feature stacking
Speech emotion recognition
Transformer

ABSTRACT

This study evaluates the effectiveness of data augmentation on 1D convolutional neural network (CNN) and transformer models for speech emotion recognition (SER) on the Ryerson audio-visual database of emotional speech and song (RAVDESS) dataset. The results show that data augmentation has a positive impact on improving emotion classification accuracy. Techniques such as noising, pitching, stretching, shifting, and speeding are applied to increase data variation and overcome class imbalance. The 1D CNN model with data augmentation achieved 94.5% accuracy, while the transformer model with data augmentation performed even better at 97.5%. This research is expected to contribute better insights for the development of accurate emotion recognition methods by using data augmentation with these models to improve classification accuracy on the RAVDESS dataset. Further research can explore larger and more diverse datasets and alternative model approaches.

Corresponding Author:

M. Anang Mukhlas
Department of Computer Science, BINUS Graduate Program, Master of Computer Science
Bina Nusantara University
Jakarta, 11480, Indonesia
Email: m.mukhlas@binus.ac.id

1. INTRODUCTION

Speech emotion recognition (SER) is the ability of computers to recognize and understand human voices and convert them into text or commands to be executed. SER is a widely used technique for efficient information sharing and communication between people and computers. It has a wide range of real-world uses in the field of human-computer interaction (HCI) [1]. This technology can be used in various fields such as developing more humane human-computer interaction systems, better speech recognition and in the field of psychology to study the relationship between sound and emotion. For example, virtual assistants that can react to users' emotions more effectively, surveillance systems that can identify drivers' levels of stress, or assistive technology for persons with disabilities that can react to their needs and emotions with greater sensitivity. The monitoring and treatment of emotional problems as well as mental health research can both benefit from emotion identification in speech in the healthcare industry.

In recent years, SER has become an interesting research material that all researchers worldwide can develop. With the development of artificial intelligence (AI), the interaction between humans and computers becomes more comfortable and how to better use AI and the development of SER is the main focus for the next generation [2]. But doing research on emotion recognition is very difficult to do because distinguishing emotions from speech features becomes unclear due to the direct influence of differences in sentences, speakers, speech styles [3]. And also, people often use multiple cues to convey their emotions

simultaneously, such as speech characteristics, language content, facial expressions, and body movements, SER is essentially a complicated multimodal task [4].

We have found the paper that is the reference for our paper using the Ryerson audio-visual database of emotional speech and song (RAVDESS) dataset and the convolutional neural network (CNN) method and a combination of five features, namely Mel-frequency cepstral coefficients (MFCC), Mel-spectrogram, chronogram, spectral contrast, and tonnetz, the best accuracy is 79.17% [5]. We use this paper because this research has the best feature stack in the CNN model. From previous research that is used as a reference [5], this study aims to compare data processed without data augmentation and those that perform data augmentation, also in this study changing the parameters of the data preprocessing process, data augmentation and feature extraction to enrich the data so that it has a lot of data and data variations, also for the model used in this study is a 1D CNN model such as the baseline of previous research to see the results of changes between the many data variations carried out today with the same model.

In this paper, we conduct a SER research scheme consisting of several steps depicted in Figure 1, preprocessing, feature extraction with feature fusion and data augmentation, and finally modeling. In the data preprocessing stage, we perform initial processing of the audio data, such as data cleaning, trimming if necessary, and setting audio parameters such as duration or a consistent sampling rate. Furthermore, in the feature extract step, the process is done by feature stacking and data augmentation as shown in Figure 1.

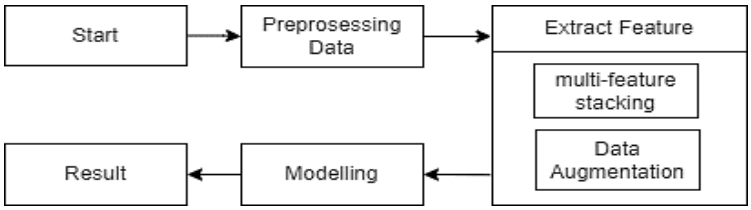

Figure 1. SER processs

Extract features here involves extracting features from the audio data using techniques such as Mel spectrogram, MFCC, chroma, and so on. We also combine these features by stacking. We also tested the effect of data augmentation, where we extracted features from augmented data and data without augmentation for comparison.

Based on Figure 1, the next stage is modeling where researchers build models, namely CNN and transformers used in this study. After building the model, the next step is to train the model using data that has previously been separated into training, validation, and testing data. The model evaluation that will be used in this research is the accuracy metric. After training the model and obtaining the results, the last step is to analyze the results obtained and compile a research report. In the report, we make a comparison of the performance of the CNN and transformer models, as well as the effect of using data augmentation or no data augmentation on the performance of the models.

2. METHOD

Preprocessing is a technique performed before extracting signal features performed on speech samples [6]. Preprocessing plays a crucial role in improving the quality and relevance of the speech data before feature extraction and model training. Several preprocessing techniques are commonly applied to ensure accurate emotion recognition. In this research, the dataset used is the RAVDESS dataset [7]. This dataset consists of 1440 male and female audio speech data files with a total of 24 professional actors recording 8 different emotions, namely happy, calm, neutral, angry, fear, disgust, shock, and sadness on English audio speech.

The dataset thus offers a diverse range of vocal expressions, allowing researchers to delve into the intricate aspects of speech analysis and emotion recognition. The initial process in this research involves categorizing the dataset into emotions and audio file paths. This information is stored in the emotion_df variable as tuples containing emotion and path. Next, feature extraction is performed on the audio data using various data augmentation techniques.

The purpose of data augmentation is to increase the amount of effective training and testing data and to avoid data deficiencies and help machine learning models to learn more useful features and to avoid data deficiencies [8]. These techniques include noising, stretching, shifting, pitching, speed up, and speed down. Noising adds noise to audio data to generate data variation. Stretching changes, the duration of the audio

data, while shifting adjusts the audio data within a specific time range. Pitching changes the pitch of the audio data, while speed up and speed down speed up or slow down the speed of the audio data.

After data augmentation, feature extraction is performed using several methods. Feature extraction in SER is very important as it helps to improve recognition accuracy and performance of speech signals [9]. The features we use in this research are MFCC, chromagram, Mel-spectrogram, spectral contrast, and tonnetz because they are the best features from previous research [4] and MFCC and Mel-spectrogram are widely used in SER [10], [11]. Spectral contrast can be defined as the decibel difference between peaks and valleys in a spectrum [12]. Spectral contrast, is an energy contrast that calculates the ratio of peak energy to valley energy in each vocal fold converted from spectrogram frames [13].

Tonnetz is a 6-dimensional pitch shape used to capture traditional harmonic relationships [9]. This feature extraction technique is used in audio signal processing to analyze the tonal centroid and features of the audio signal to learn the features that are being extracted from the audio file [14]. Chromograms are signal features altered by spectra created to capture harmony and melodic characteristics by using scales [13]. The chromogram feature becomes a tool capable of analyzing music data in representing the tonal content of music audio signals in a condensed form [15].

Mel-spectograms are used to identify and track timbre fluctuations in sound files and tend to be poor at distinguishing pitch and harmony classes [5]. The process to obtain a Mel-spectogram is that the spectrogram is derived by short-term Fourier transform and then the resulting spectrogram is transformed into the human perception scale [16]. In order to extract spectrogram characteristics from an audio signal, it must first be divided into brief overlapping windowing, transformed into the frequency domain using the Fourier transform, and then generated into an envelope spectrogram using a Mel filter bank (see Figure 2) [17]. MFCC, which stands for Mel frequency spectrum per frame, is a type of signal spectrum that can be obtained collectively [13].

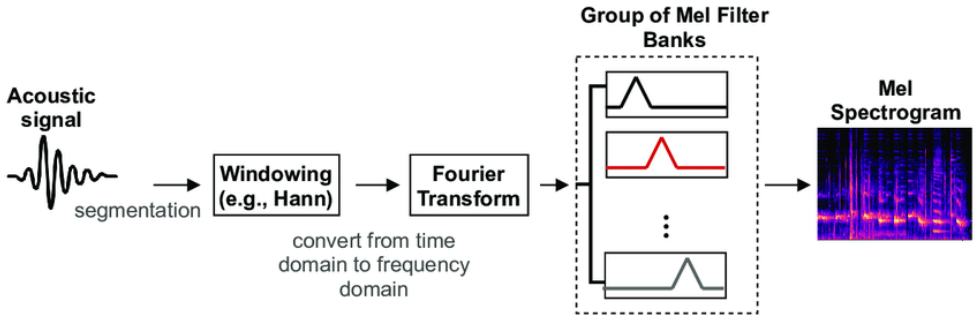

Figure 2. Process of extracting Mel-spectogram

MFCC allows for the representation of signals more in line with human perception through mapping into the Mel scale, which is an adaptation of the Hertz scale for frequencies in the human sense of hearing [18]. These MFCCs can be used for speech emotion analysis and classification [19]. The parameters used in the feature extraction process are a sample rate of 22050, an FFT frame size (n_fft) of 2048, and a frame shift (hop_length) of 512. The extracted features are then stored in variables X and Y. Variable X is an array containing audio features with a dimension of 211.

This dimension is obtained by summing up the number of features from each feature extraction method: spectral contrast (7 features), Tonnetz (6 features), chromagram (12 features), Mel spectrogram (128 features), and MFCC (58 features). Overall, there are 211 feature values represented in variable X. On the other hand, variable Y is an array containing emotion labels. This dataset includes 8 emotion labels that are used to describe the voice expression in the audio. Each emotion label has a corresponding numerical code: 01 for "neutral" emotion, 02 for "calm" emotion, 03 for "happy" emotion, 04 for "sad" emotion, 05 for "angry" emotion, 06 for "fear" emotion, 07 for "disgust" emotion, and 08 for "surprised" emotion.

After performing feature extraction, the next step is to combine all the features that have different sizes used into one array. Each feature will be compressed into a unified array dimension by taking the mean value and then stacking it according to the array queue that was created before [4]. From the results of Choi *et al*. [20] the effect of using the stacking feature can increase performance accuracy by 2% because the stacking feature can emphasize the required respiration information and diversify bandwidth.

After obtaining the X and Y values, researchers divided the data with a proportion of 70% for training, 10% for validation, and 20% for testing with a random state of 42. After the data was divided for

training, validation, and testing, the next step was to standardize the data to ensure the data had the same mean and standard deviation values. The dimensions of variable X in training, validation, and testing were expanded. The data is then entered into the model. The first model used is the CNN model as a comparison with previous research with the same model.

CNN is a deep learning algorithm that has high performance in image classification and consists of several successive layers that differ in parameter selection. In the context of SER, CNN is used to automatically extract acoustic features from speech spectrograms or other representations. The benefit of applying CNN is to extract features by generating detailed structural feature maps from time series and provide comparable performance to self-constructed feature matching models [21]. This time, the 1D CNN model will use aggregated and unaggregated data to see the changes that occur, and Figure 3 shows an example of a simple one-dimensional CNN architecture.

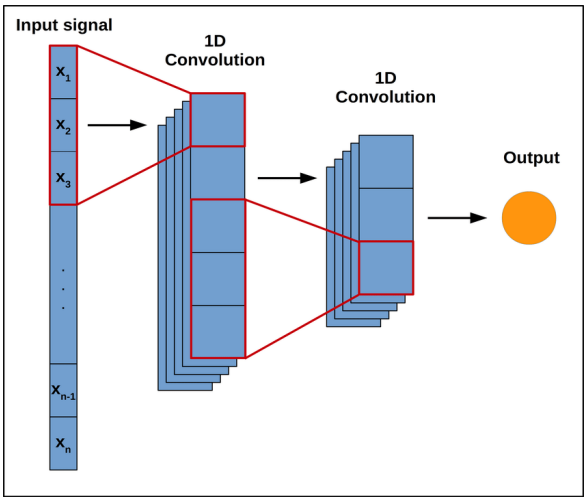

Figure 3. Architecture of 1D CNN

Figure 3 shows the convolution stages each with a group of learnable convolution filters and grouping operations. By performing heavy convolution with the input and applying a non-linear activation function, this convolution filter works to extract high-level features. After that, these features are fed into a clustering operation, which increases the spatial size of the features extracted by the convolution filters while emphasizing the dominant features learned by each filter. As the input is processed through convolution stages, the network acquires characteristics that are more specific to the problem at hand.

Next, researchers built a 1D CNN model. This model uses several 1D convolution layers to extract features from audio data. First, a Conv1D layer with 32 filters and a Kernel size of 3 was applied with a rectified linear unit (ReLU) activation function. This was followed by a MaxPooling1D layer with a pooling size of 2 to reduce the dimensionality of the data.

This process is repeated with more complex Conv1D and MaxPooling1D layers, with the number of filters increasing to 64 and 128. The output of the convolution layer is flattened into a one-dimensional representation using the flatten layer. This is done to allow the output to be processed by the next dense layer. A dense layer with 256 units and ReLU activation function is added to process the extracted features.

Dropout with a rate of 0.5 is applied to reduce overfitting. Finally, a dense layer with 8 units and a softmax activation function is used to generate a classification output corresponding to the number of emotion labels. Then we also use the transformer model as a comparison model for CNN. In SER, the Transformer model is a neural network that is better at capturing temporal context than recurrent neural networks (RNNs) at capturing context-dependent emotions.

As a result, SER researchers were inspired to use transformers in their research [22]. For the transformer model in this study, we used typical transformer layers such as multi-head attention and feed forward network. The multi-head attention layer is used to extract a more robust representation of audio features by considering the interactions among different parts. The feed forward network layer applies linear and non-linear operations to process the previously obtained feature representation.

To build the transformer model, a transformerlayer object is initialized, which includes the multi-head attention and feed forward network layers. This object is then used as a layer in the transformer model. The model also includes a flatten layer to convert the feature representation into a one-dimensional form and

a dense layer with a softmax activation function to generate a classification output corresponding to the number of emotion labels. The parameters used in the transformer model include 512 for the feature representation dimension (d_model), 4 for the number of heads in the multi-head attention layer, 256 for the number of units in the layer.

Feed forward network layer, 0.2 for the dropout rate, and 8 for the number of emotion classes. Next, the 1D CNN model and transformer model were compiled using the Adam optimizer, the categorical_crossentropy loss function, and the accuracy metric. ReduceLROnPlateau and EarlyStopping callbacks were used to adjust learning and prevent overfitting. The number of epochs used was 100, and the batch size was set to 64.

3. RESULTS AND DISCUSSION

Our study was conducted in 4 experiments using 2 models, namely 1D CNN and transformer, with and without data augmentation. We used the RAVDESS dataset for this study. We found that the use of data augmentation on the RAVDESS dataset was highly effective due to the presence of imbalanced class distribution.

In this study we used five data augmentation techniques applied to the RAVDESS dataset, namely noising, pitching, stretching, shifting, and speeding. These augmentation techniques successfully increased the training sample size from 1440 to 10080, provided better data variation, and helped overcome the imbalance problem. After data augmentation, we used feature extraction using several methods namely spectral contrast, Tonnetz, chromagram, Mel spectrogram, and MFCC.

In Table 1, we conducted the first experiment using the 1D CNN model without augmentation and achieved an accuracy of 72.4%. Furthermore, we used the transformer model without augmentation and obtained an accuracy of 74.0%. However, we also found that using data augmentation significantly improved the performance of the model. In the third experiment, we implemented data augmentation on a 1D CNN model and achieved 94.5% accuracy. This shows that data augmentation is very effective in improving the performance of the 1D CNN model.

Table 1. The comparison table of experiment

Previous work	Feature order	Database	Method	Accuracy (%)
Tanoko and Zahra [5]	MFCC	RAVDESS	ResNet transformer–encoder CNN	80.8
Er [23]	MFCC, RMS, cromag, spectral, and spectrogram	RAVDESS	ResNet101 (augmentation)	79.4
Han *et al.* [24]	Spectral contrast, MFCC, Tonnetz, chromagram, and Mel-spectogram	RAVDESS	1D CNN (augmentation)	79.1
Bhangale and Kothandaraman [25]	Multiple acoustic (spectral, time domain, and voice quality)	RAVDESS	1D DCNN	94.1
Our model	Spectral contrast, MFCC, Tonnetz, Chromagram, and Mel-spectogram	RAVDESS	1D CNN	72.4
			Transformer	74.0
			1D CNN (augmentation)	94.5
			Transformer (augmentation)	97.5

Our last experiment involved using the transformer model with data augmentation, and we achieved an excellent accuracy of 97.5%. Comparison with other studies also provided interesting results which are shown in Table 1. In the study by Tanoko and Zahra [5], the use of multi-features with a 1D CNN model with augmentation resulted in 79.1% accuracy.

However, in our study, the 1D CNN model with data augmentation and multi-feature achieved higher accuracy, demonstrating the effectiveness of data augmentation. Comparison with other studies showed that our implemented 1D CNN model with data augmentation can compete with 1D DCNN and transformer models. In the study by Bhangale and Kothandaraman [25], the 1D DCNN model achieved an accuracy of 94.1%, while in our research, the 1D CNN model with data augmentation achieved similar accuracy. This suggests that with appropriate data augmentation, the 1D CNN model can perform on par with other more complex models.

Overall, our research results demonstrate that the use of data augmentation in conjunction with the 1D CNN and transformer models can improve emotion classification accuracy on the RAVDESS dataset. Particularly, the transformer model exhibited remarkably high accuracy, surpassing even the performance of the 1D CNN model with data augmentation. These findings contribute valuable insights to the development of more accurate and reliable emotion recognition methods.

4. CONCLUSION

This research evaluates the use of data augmentation in 1D CNN and transformer models for SER on the RAVDESS dataset. The results show that data augmentation has a positive impact on improving emotion classification accuracy. The use of augmentation techniques such as noising, pitching, stretching, shifting, and speeding successfully increases the variety of training data and overcomes the problem of class imbalance in the dataset. In this experiment we have tested the 1D CNN model with data augmentation achieved an accuracy of 94.5%, while the transformer model with data augmentation had a higher accuracy of 97.5%. These results show that the transformer model has better performance in SER on the RAVDESS dataset. Comparisons from previous studies show that our 1D CNN model with data augmentation is able to compete with more complex models such as 1D DCNN and transformers. This study outperformed the accuracy achieved in previous studies, showing that proper use of data augmentation can result in significant performance improvements.

The results of this research can provide valuable insights for the development of more accurate and reliable emotion recognition methods. The use of data augmentation in combination with 1D CNN and transformer models can improve the accuracy of emotion classification on the RAVDESS dataset. These findings can be applied in various applications that require emotion analysis in speech. We recommend the use of data augmentation as an effective strategy in improving the performance of emotion recognition models. Furthermore, this research can be extended by using larger and more diverse datasets for more comprehensive validation and testing. Other model development such as model combination or the use of other methods can also be an interesting research area in the future.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Kaggle for providing the invaluable dataset that served as the primary foundation of this research. We extend our appreciation to Binus University for their financial support in the publication of this study and to the entire journal staff for their assistance, guidance, and support throughout the editorial and publication process. The contributions and collaboration of our fellow researchers are also greatly valued; they have provided invaluable insights and knowledge throughout the course of this research.

REFERENCES

[1] Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020, doi: 10.1109/ACCESS.2020.2990405.

[2] P. Jiang, H. Fu, H. Tao, P. Lei, and L. Zhao, "Parallelized Convolutional Recurrent Neural Network With Spectral Features for Speech Emotion Recognition," *IEEE Access*, vol. 7, pp. 90368–90377, 2019, doi: 10.1109/ACCESS.2019.2927384.

[3] A. B. Ingale and D. S. Chaudhari, "Speech Emotion Recognition," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 2, no. 1, pp. 235–238, 2012.

[4] A. S. Haq, M. Nasrun, C. Setianingsih, and M. A. Murti, "Speech recognition implementation using MFCC and DTW algorithm for home automation," *Proceeding of the Electrical Engineering Computer Science and Informatics*, vol. 7, no. 2, pp. 78–85, 2020.

[5] 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.

[6] N. Kurpukdee, S. Kasuriya, V. Chunwijitra, C. Wutiwiwatchai, and P. Lamsrichan, "A study of support vector machines for emotional speech recognition," in *2017 8th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)*, IEEE, May 2017, pp. 1–6, doi: 10.1109/ICTEmSys.2017.7958773.

[7] 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.

[8] S.-T. Pan and H.-J. Wu, "Performance Improvement of Speech Emotion Recognition Systems by Combining 1D CNN and LSTM with Data Augmentation," *Electronics*, vol. 12, no. 11, p. 2436, May 2023, doi: 10.3390/electronics12112436.

[9] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019, doi: 10.1109/ACCESS.2019.2938007.

[10] 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)*, IEEE, Apr. 2019, pp. 1–6, doi: 10.1109/RADIOELEK.2019.8733432.

[11] 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)*, IEEE, Mar. 2017, pp. 2257–2260, doi: 10.1109/WiSPNET.2017.8300161.

[12] 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.

[13] Z. Dair, R. Donovan, and R. O'Reilly, "Linguistic and gender variation in speech emotion recognition using spectral features," *arXiv preprint arXiv:2112.09596*, 2021.

[14] E. J. Humphrey, T. Cho, and J. P. Bello, "Learning a robust Tonnetz-space transform for automatic chord recognition," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Mar. 2012, pp. 453–456, doi: 10.1109/ICASSP.2012.6287914.

[15] A. K. Shah, M. Kattel, A. Nepal, and D. Shrestha, "Chroma feature extraction," *Chroma Feature Extraction using Fourier*

Transform, 2019.
[16] S. Zhilibayev, "Real-time Speech Emotion Recognition (RSER) in Wireless Multimedia Sensor Networks," 2021.
[17] M. Habib, M. Faris, R. Qaddoura, M. Alomari, A. Alomari, and H. Faris, "Toward an Automatic Quality Assessment of Voice-Based Telemedicine Consultations: A Deep Learning Approach," *Sensors*, vol. 21, no. 9, p. 3279, May 2021, doi: 10.3390/s21093279.
[18] Ms. Shambhavi. S. Sheerur and Dr. V. N. Nitnaware, "Emotion Speech Recognition using MFCC and SVM," *International Journal of Engineering Research and*, vol. V4, no. 06, Jun. 2015, doi: 10.17577/IJERTV4IS060932.
[19] A. Mahmood and K. Ö. S. E. Utku, "Speech recognition based on convolutional neural networks and MFCC algorithm," *Advances in Artificial Intelligence Research*, vol. 1, no. 1, pp. 6–12, 2021.
[20] Y. Choi, H. Choi, H. Lee, S. Lee, and H. Lee, "Lightweight Skip Connections With Efficient Feature Stacking for Respiratory Sound Classification," *IEEE Access*, vol. 10, pp. 53027–53042, 2022, doi: 10.1109/ACCESS.2022.3174678.
[21] A. Shenfield and M. Howarth, "A Novel Deep Learning Model for the Detection and Identification of Rolling Element-Bearing Faults," *Sensors*, vol. 20, no. 18, p. 5112, Sep. 2020, doi: 10.3390/s20185112.
[22] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. Schuller, "Survey of Deep Representation Learning for Speech Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1634–1654, Apr. 2023, doi: 10.1109/TAFFC.2021.3114365.
[23] M. B. Er, "A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features," *IEEE Access*, vol. 8, pp. 221640–221653, 2020, doi: 10.1109/ACCESS.2020.3043201.
[24] S. Han, F. Leng, and Z. Jin, "Speech Emotion Recognition with a ResNet-CNN-Transformer Parallel Neural Network," in *2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*, IEEE, May 2021, pp. 803–807, doi: 10.1109/CISCE52179.2021.9445906.
[25] K. Bhangale and M. Kothandaraman, "Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network," *Electronics*, vol. 12, no. 4, p. 839, Feb. 2023, doi: 10.3390/electronics12040839.

BIOGRAPHIES OF AUTHORS

Khasyi Al Mukarram is an undergraduate and postgraduate student in the Bina Nusantara University master track program who is currently interested in speech technology such as speech recognition, speech translation, speaker recognition, and natural language processing (NLP). He is also interested in studying other fields such as deep learning and virtual reality (VR). He can be contacted at email: Khasyi.mukarram@binus.ac.id.

M. Anang Mukhlas is a student of Computer Science at Bina Nusantara University, Indonesia, since 2019. He studied to earn a bachelor's degree as well as a master's degree concurrently in the master track program at Bina Nusantara University. His research is interested in speech recognition in the field of speech technology deep learning and natural language processing (NLP); additionally, he is also interested in the fields of image processing and machine learning. He can be contacted at email: m.mukhlas@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.