

# Speech emotion recognition using 2D-convolutional neural network

Fauzivy Reggiswarashari, Sari Widya Sihwi

Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret, Surakarta, Indonesia

## Article Info

### Article history:

Received Aug 31, 2021

Revised May 28, 2022

Accepted Jun 26, 2022

### Keywords:

2D-CNN

Audio segmentation

Mel frequency cepstral coefficient

Speech emotion recognition

## ABSTRACT

This research proposes a speech emotion recognition model to predict human emotions using the convolutional neural network (CNN) by learning segmented audio of specific emotions. Speech emotion recognition utilizes the extracted features of audio waves to learn speech emotion characteristics; one of them is mel frequency cepstral coefficient (MFCC). Dataset takes a vital role to obtain valuable results in model learning. Hence this research provides the leverage of dataset combination implementation. The model learns a combined dataset with audio segmentation and zero padding using 2D-CNN. Audio segmentation and zero padding equalize the extracted audio features to learn the characteristics. The model results in 83.69% accuracy to predict seven emotions: neutral, happy, sad, angry, fear, disgust, and surprise from the combined dataset with the segmentation of the audio files.

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



## Corresponding Author:

Sari Widya Sihwi

Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret

Jl. Ir. Sutami No 36A, Surakarta, Central Java, 57126, Indonesia

Email: sariwidya@staff.uns.ac.id

## 1. INTRODUCTION

Artificial intelligence (AI) studies show enormous beneficial growth in most aspects of the product industry. One of the studies in AI that gather the researcher's attention is affective computing with all its humanitarian approaches. Affective computing recognizes, processes, and produces human feelings using a computational approach [1]. It reaches many fields in the industry, such as medic, psychotherapy, marketing, and advertising [2]. The implementation of emotion detection expands facial expression recognition, language recognition, empathy giving, and other functionalities closer to how humans function. In addition, emotion recognition widens the technology growth of affective computing itself. Emotion recognition is human feelings recognition of many data sources, from texts, facial expressions, voices, gestures, and behaviors [3]–[5].

Researchers face two main focuses to solve in speech emotion recognition. They are the extracted feature selection and model selection. Speech emotion recognition is an emotion recognition of human speech using computation models in affective computing and develops forensic, security, and biometric fields [6]. Besides its lexical contents, human speech expresses other characteristics: information of age, gender, language, and emotion [6], [7]. The audio contains many features to extract into computational form by transforming audio to a graph of arrays. The extraction of speech waveform is categorized into four: prosodic, spectral, audio quality, and Teager energy operator features [8]. Extracted audio features are mel frequency cepstral coefficient (MFCC), spectrogram, zero crossing rate (ZCR), Teager energy operator (TEO), harmonic to noise rate (HNR) [9]–[12]. In emotion recognition, MFCC is a cepstral domain feature used the most for research [12].

For decades, speech emotion recognition has been done and delivered different results and developments. Extracted audio feature selection takes effect on emotion prediction results. Based on prior research, researchers often use MFCC, and the feature results in better performance [10], [13]–[15]. To maximize the usage of those features, segmentation on speech data may enhance the accuracy of emotion prediction, like what Yeh *et al.* [16]. They produced a series of segments by parting a segment of existing speech from a long audio track. A frame-based segmentation applied well to a track of singular emotion speech, but voiced segmentation is better to apply in a long speech.

Learning the emotions can be done by using models of deep learning. The growth of emotion recognition systems is also affected by learning models, like deep learning, which can recognize image structures of voices and facial expressions [1]. A deep learning advantage for emotion recognition is the automation of feature selection that contains important emotion attributes of the data source, especially audio [17].

Researchers use the classification models to classify emotions from extracted audio features. Some of them are convolutional neural network (CNN), recurrent neural network (RNN), and support vector machine (SVM) [10], [14], [17] among those classification models, CNN is confident in processing pictures and applying them to speech emotion recognition since the extracted audio features form can be a picture. Prior research of speech emotion recognition shows that CNN produces higher accuracy than RNN and SVM models [14], [18]. This condition allows us to choose CNN as the platform for the speech emotion recognition model. CNN architecture consists of an input layer, convolutional layer, pooling layer, fully-connected layer, and output layer sequentially [19]. The essential operation of CNN is in the convolution and pooling process; the convolution process uses filters to extract feature maps of datasets where beneficial information is maintained, pooling or subsampling process reduces feature maps dimension [20] and could benefit in reverberant conditions [21]. CNN shows its confidence in learning emotions for speech emotion recognition with high accuracy [13]–[15], [22].

Speech emotion recognition research uses many audio data from datasets that vary in languages, duration, and emotions. Most speech emotion recognition research uses datasets equipped with emotion labels. Some of the datasets used by researchers are Ryerson audio-visual database of emotional speech and song (RAVDESS), surrey audio-visual expressed emotion (SAVEE), The interactive emotional dyadic motion capture (IEMOCAP), Toronto emotional speech set (TESS), CASIA, dan EMO [9], [11], [14], [15], [22], [23]. The work to combine datasets is expanded by de Pinto *et al.* [15] to gain a model with better performance. They combined the RAVDESS speech and TESS datasets to minimize their prior work's overfitting condition using one-dimensional convolutional network; hence, it reached a better model than the particular dataset from their prior work. Overfitting is the network's inability to learn data effectively for various reasons, such as problems in learning noise in the data train, imbalanced variance and bias, and ineffective architecture [24]. The combination of datasets may deliver more significant data quantities to enhance the model performance on learning more training sets.

Therefore, compared to prior works, we propose a different approach: implementing a two-dimensional convolutional neural network (2D-CNN) that utilizes MFCC features and audio segmentation with a larger dataset. We combine RAVDESS, SAVEE, and TESS datasets. The three datasets have similar pitch-based emotions and similar audio lengths with seven same emotions: neutral, sad, happy, angry, disgust, fear, and surprise. We deliver a 2D-CNN by adding MFCC as an additional dimension in our classification model.

## 2. METHOD

### 2.1. Dataset

First of all, we find all available labelled speech audio datasets on the internet. The chosen datasets are the ones that fulfil the needs of experiments such as having similar emotions, similar track duration, and the works of literature that used the datasets. In this research, we picked several high-quality datasets with useful similarities for learning since it causes the possibility of accuracy reduction of deep learning algorithms to be minimized [25].

The collected datasets are RAVDESS, SAVEE, and TESS, which have exactly one emotion in each audio track and range from one to five seconds of duration in each track. RAVDESS is a database of audio speech and song in English with eight validated emotions [26]. RAVDESS has 24 actors who expressed neutral, calm, happy, sad, angry, fear, surprise, and disgust in a total of 1,440 speech data. SAVEE is a recorded visual and audio dataset from four researchers at the University of Surrey [2]. SAVEE has seven emotions: anger, disgust, fear, happiness, sadness, and surprise, expressed in English with 480 tracks. TESS (Toronto emotional speech set) is a set of audio from two actresses who expressed seven emotions: anger, disgust, fear, happiness, sad, neutral, and pleasant surprise [27]. The TESS dataset consists of 2,800 audio files in English.

## 2.2. Preprocessing

The preprocessing data step consists of segmentation, audio feature extraction, and zero padding of MFCC. We use Python in the Jupyter Notebook equipped with Librosa, Numpy, operating systems (OS), and JSON libraries to do this preprocessing step. Two types of segmentation models were carried out in the experiment, namely linear segmentation and overlap segmentation. Linear segmentation cuts voice recordings sequentially, without any segments having data in common. Overlap segmentation cuts voice recordings by making data slices that intersect in each segment. The sample MFCC data on RAVDESS Figure 1(a) after segmentation is Figure 1(b) for linear segmentation and Figure 1(c) for overlap segmentation. Next, to avoid leaving important parts behind segmentation and help each segment to have the same length, we use zero padding on imperfect segments. Zero padding helps the imperfect segments to reach the target length by padding the transposed MFCC array of the segment with zeros. The MFCC array is transposed first to be processed in classification later. Figure 1(d) shows the applied zero padding on an imperfect segment.

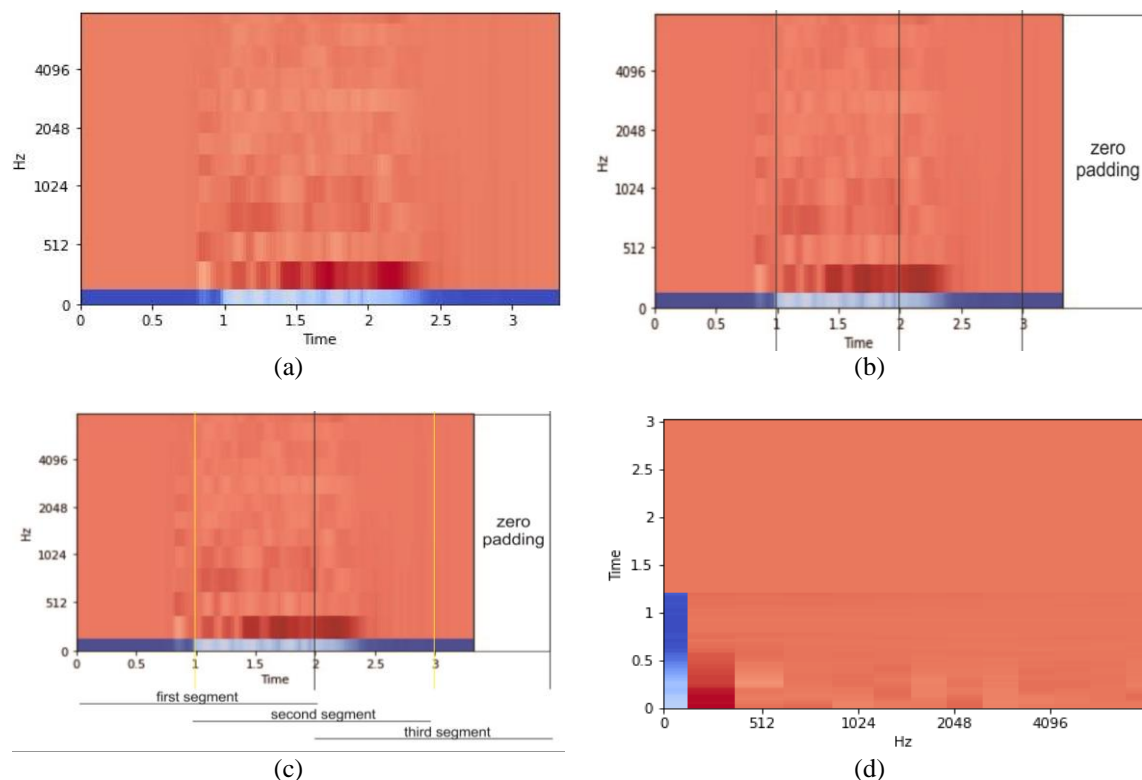


Figure 1. Transformation of MFCC of audio sample in RAVDESS, from (a) an original MFCC data before preprocessing steps, then after preprocessing steps; (b) linear segmentation of one-second duration; (c) overlap segmentation of two-second duration; and (d) zero padding on a segment.

## 2.3. Emotion classification experiments

The processing data step consists of the emotion classification by experimenting on the model architecture and datasets. In this step, we use Python in the Jupyter Notebook equipped with TensorFlow, Sklearn, Numpy, JSON, dan Matplotlib libraries. The JSON file of the dataset will be loaded then grouped into sets such as train set, test set, and validation set. The MFCC array is then formed into a three-dimensional array by adding a new axis as depth. The three-dimensional shape represents the length of MFCC, the number of MFCC, and the visual color of the picture. For instance, one represents grayscale, and three represents red, green, blue (RGB) color.

The CNN architecture is built using the Keras library from TensorFlow. It consists of essential CNN layers: convolutional layers, pooling layers, and fully connected layers, which are stacked sequentially in a model. We use a block of convolutional layers consisting of a 2D convolutional layer, a 2D max-pooling layer, and a batch normalization layer. The convolutional layer is activated using the ReLu function. The next layer is a flattening layer for deducting the shape into one vector data; hence the data can be classified. After

the flattening layer, a fully connected layer is implemented to classify. The output layer is activated using the SoftMax function. The learning process uses the Adam optimizer from Keras with the loss calculated by sparse categorical cross-entropy.

We do some experiments for model enhancement to avoid overfitting, including model performance improvements. There are three aspects in our experiments: CNN network, learning process, and dataset. In the experiments of CNN network, we focus on layers, filters, and also regularization. In the aspect of the learning process, the independent parameters used are the dataset allocation ratio, batch size, and epochs. Furthermore, as the other aspects, we also have some experiments in the dataset aspect, which are segmentation and duration. In the end, all experiments from these three aspects will produce the best-experimental model.

## 2.4. Evaluation

The evaluation step will rate the performance of the model. The best-experimented model from processing data steps is used to predict emotions from a prepared dataset sample. The model will show the prediction of some samples' emotions. Then, we evaluate the model's performance by measuring its precision, recall, and F1-score. Next, the model is compared with other models to see the quality of its performance.

## 3. RESULTS AND DISCUSSION

### 3.1. Preprocessing data

Preprocessing step includes segmentation, feature extraction, and zero padding. Each track has a different duration length from one to another. As we use a frame-based segmentation, we first decide the length of the segment by analyzing the length of the tracks. We consider the possible length of the segment to be between one to three seconds. Then, a duration length experiment will evaluate the best length for the segment.

We use the two types of frame-based segmentation model, namely linear segmentation and overlap segmentation. After the segmentation, the MFCC feature of each segment is extracted. We use 13 MFCCs; therefore, the shape of extracted MFCC is (13, n). For classification, we transpose the array to the shape of (n, 13). Because the segmentation may have leftover segments with less than the desired length, zero padding is applied to help the segments reach the desired length. We pad the MFCC array using zeros to the desired shape. In audio data, this means adding silent audio to the segment. For example, for a three-second segment with a shape of (130, 13), on a (52, 13) segment, we add (78, 13) zeros to it, so it reaches the desired shape (130, 13). After the entire segments are ready, we save them into JSON list data. Table 1 presents the number of segments in each dataset.

Table 1. Preprocessing data results

| Duration | Dataset | Emotions |      |      |      |     |      |      | Total |
|----------|---------|----------|------|------|------|-----|------|------|-------|
|          |         | N        | H    | Sad  | A    | F   | D    | Sur  |       |
| 2s       | RAVDESS | 193      | 401  | 408  | 450  | 397 | 461  | 389  | 2699  |
|          | SAVEE   | 270      | 142  | 166  | 141  | 140 | 148  | 141  | 1148  |
|          | TESS    | 663      | 567  | 798  | 564  | 107 | 800  | 583  | 4082  |
|          | Total   | 1126     | 1110 | 1372 | 1155 | 744 | 1409 | 1113 | 7929  |
| 3s       | RAVDESS | 192      | 384  | 384  | 384  | 384 | 384  | 383  | 2495  |
|          | SAVEE   | 215      | 105  | 119  | 107  | 107 | 113  | 106  | 872   |
|          | TESS    | 400      | 400  | 400  | 400  | 400 | 400  | 400  | 2800  |
|          | Total   | 807      | 889  | 903  | 891  | 891 | 897  | 889  | 6167  |

### 3.2. Processing data

Processing data is the process of training a model to classify emotions. The first thing to do in processing data is load data. We pick the prepared dataset in loading data and then convert the MFCC arrays into a four-dimensional array to value its depth of color dimension. The shape of MFCC arrays is now (m, n, 13, 1) with m as the number of segments and n as the MFCC length. MFCC arrays are the classification data. Emotion labels are the classification targets. The data are then grouped into three categories: train set, test set, and validation set with a ratio of 5:3:2 to 8:1:1.

The proposed architecture of CNN Network consists of 2D convolutional blocks, a flattening layer, and a fully connected layer. The model is affected by the parameters we choose in model optimization. The model optimization applied in this research is shown in Table 2.

Each experiment is held under the same circumstances to check different effects from assigned parameters. The result shows unexpected overfitting conditions in several experiments. Through the experiments, we could define better parameters to minimize overfitting conditions. For example, in dropout

experiments, implementing dropout reduces the overfitting because dropout updates the outgoing edges of neurons to zero. Implementing regularizers also minimize overfitting since they ease model learning complexity. In our case, the primary reason overfitting happens is the small quantity dataset. The epochs experiment shows that a small quantity dataset would result in overfitting as adding the number of epochs.

Table 2. Experimental table

| Aspect                     | Experiments                          | Parameters                            | Results (Avg of Acc) |
|----------------------------|--------------------------------------|---------------------------------------|----------------------|
| CNN Network                | Layers experiments                   | <b>Two convolutional layer blocks</b> | <b>66.05%</b>        |
|                            |                                      | Three convolutional layer blocks      | 63.67%               |
|                            | Filter shape experiments             | 3×3 and 3×3                           | 62.34%               |
|                            |                                      | <b>3×3 and 2×2</b>                    | <b>67.96%</b>        |
|                            |                                      | 2×2 and 3×3                           | 63.40%               |
|                            |                                      | 2×2 and 2×2                           | 66.58%               |
|                            | Filter amount experiments            | 32 and 32                             | 63.15%               |
|                            |                                      | 32 and 64                             | 67.05%               |
|                            |                                      | 64 and 32                             | 68.01%               |
|                            |                                      | <b>64 and 64</b>                      | <b>69.99%</b>        |
|                            | Dropout layer experiments            | No dropout                            | 71.88% (Overfitting) |
|                            |                                      | One dropout 0.5                       | 73.99% (Overfitting) |
| <b>Two dropouts 0.5</b>    |                                      | <b>66.69%</b>                         |                      |
| Regularization experiments | No regularizer                       | 67.45% (Overfitting)                  |                      |
|                            | <b>With regularizer l2</b>           | <b>66.94%</b>                         |                      |
| Learning Process           | Dataset ratio allocation experiments | 5:3:2                                 | 69.26%               |
|                            |                                      | 7:2:1                                 | 67.67%               |
|                            |                                      | <b>8:1:1</b>                          | <b>70.61%</b>        |
|                            | Batch size experiments               | <b>44</b>                             | <b>71.19%</b>        |
|                            |                                      | 72                                    | 68.29%               |
|                            |                                      | 100                                   | 68.83%               |
| Epochs experiments         | <b>30</b>                            | <b>72.45%</b>                         |                      |
|                            | 70                                   | 73.91% (Overfitting)                  |                      |
|                            | 100                                  | 74.12% (Overfitting)                  |                      |
| Dataset                    | Duration experiments                 | 2 secs                                | 66.93%               |
|                            |                                      | <b>3 secs</b>                         | <b>67.18%</b>        |
|                            | Segmentation experiments             | Linear segmentation                   | 63.71%               |
|                            |                                      | <b>Overlap segmentation</b>           | <b>66.59%</b>        |

The best model is formed by all those experimented parameters, which results in the highest accuracy. As shown in Table 2, the best model has two convolutional layer blocks, two dropout layers, a flattening layer, and a fully connected layer. In addition, it has filters with 3×3 and 2×2 shapes, the number of filters of 64 and 64, and regularizer L2 for CNN architecture. The dataset used to learn best combines three datasets (RAVDDESS, SAVEE, and TESS) with overlap segmentation of three-second segments. For learning the model, we use 44 batch sizes, 30 epochs, and a dataset ratio of 8:1:1. Here is the model summary presented in Figure 2 and the model performance in Figure 3.

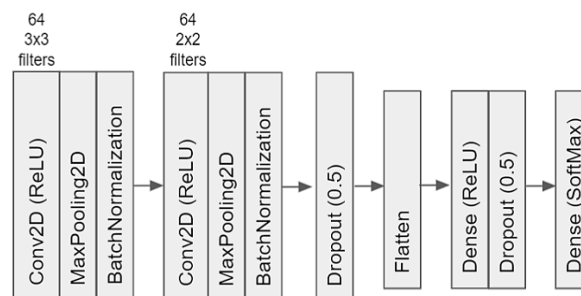


Figure 2. Model summary

### 3.3. Evaluation

In the evaluation step, firstly, we calculate the precision score, recall score, and F1-score of the model to know its performance. The dataset used is a combination of three datasets with overlap segmentation of three seconds duration. The model accuracy from the evaluation of combined datasets is

83.69%. It reaches higher accuracy because the prediction includes the same data for training. Next, we make predictions of 500 samples from the dataset containing 61 neutral, 72 happy, 74 sad, 72 angry, 74 fear, 75 disgust, and 72 surprised. From the confusion matrix, we can see the correct and false predictions. The precision, recall, and F1-score are then calculated from the correct and false predictions. We can see the scores of each emotion in Table 3. Overall, the model has an average precision score of 73.65%, a recall of 73.77%, and an F1-score of 73.63% on 500 data samples.

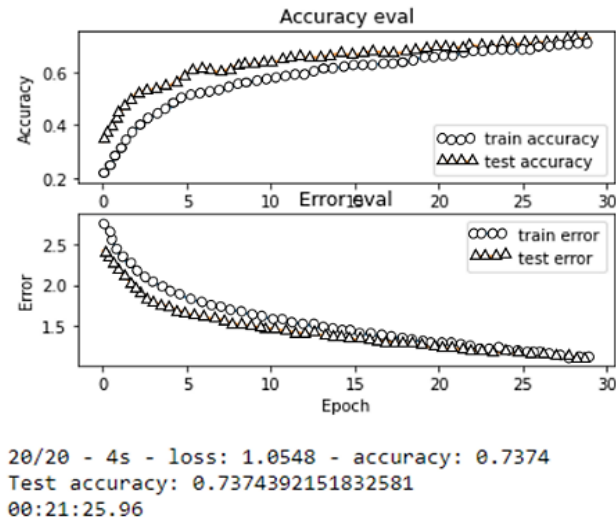


Figure 3. Model performance

Table 3. Precision, recall, and F1-score from the model

| Emotions        | Precision | Recall | F1-score | Data Count |
|-----------------|-----------|--------|----------|------------|
| Neutral         | 0.7246    | 0.8197 | 0.7692   | 61         |
| Happy           | 0.6800    | 0.7083 | 0.6939   | 72         |
| Sad             | 0.6944    | 0.6757 | 0.6849   | 74         |
| Angry           | 0.8088    | 0.7639 | 0.7857   | 72         |
| Fear            | 0.7794    | 0.7162 | 0.7465   | 74         |
| Disgust         | 0.7922    | 0.8133 | 0.8026   | 75         |
| Surprise        | 0.6761    | 0.6667 | 0.6713   | 72         |
| <b>Accuracy</b> |           |        | 0.7360   | 500        |
| <b>Average</b>  | 0.7365    | 0.7377 | 0.7363   | 500        |

After that, we evaluate the performance of the dataset combination by comparing it to each dataset. The model learns each dataset using the same test set. The comparison results in the accuracy of 40.19%, 17.99%, 53.65%, and 69.21% for the dataset RAVDESS, SAVEE, TESS, and the combination of three of them, respectively. The combination result shows higher accuracy than any individual dataset, meaning that a dataset combination could raise the model accuracy in predicting emotions.

We also compare our model to other models to see which one performs better. We experimented with the chosen combined dataset. The models to compare are CNN 1-D and RNN long short-term memory (RNN-LSTM). The CNN 1-D model uses the architecture of Pinto *et al.* [15]. The architecture of RNN-LSTM model is two LSTM layers, each followed by a dropout layer and one dense layer. The models are experimented with using similar learning process configurations on the same test set. The comparison results in 69.21%, 51.05%, and 58.02% accuracy for CNN 2-D, CNN 1-D, and RNN-LSTM, respectively. From the comparison, we can see that our proposed model, the CNN 2-D, works better than the other two.

#### 4. CONCLUSION

This research implements a 2D-CNN model to predict emotions from speech using RAVDESS, SAVEE, and TESS datasets. The extracted audio feature is mel frequency cepstral coefficient (MFCC). A frame-based segmentation is done to the audio files to get the same audio length. Each segment with less length will be padded with zeros on its MFCC array.

The optimum dataset is chosen by segmentation and combination experiments. The CNN architecture is built using experiments for model optimization, such as layers, filters, and regularizers. In addition, experiments of batch size, epochs, and dataset ratio composition optimize the learning process. The proposed model reaches 83.69% accuracy on three datasets using overlap segmentation for three-second segments. In addition, the model can classify seven emotions: neutral, happy, sad, angry, fear, disgust, and surprise. The developments for speech emotion recognition may vary. However, this research recommends the robustness of deep learning models using data augmentation techniques, building a more effective classification model, and using more precise audio features to recognize the emotion better.

## ACKNOWLEDGEMENTS

Funding was supported by Group Research Grant from non-APBN fund Universitas Sebelas Maret 2021 No: 260/UN27.22/HK.07.00/2021 for Intelligent Systems and Humanized Computing (ISHC) Research Group.




## REFERENCES

- [1] W. Dai, D. Han, Y. Dai, and D. Xu, "Emotion recognition and affective computing on vocal social media," *Information & Management*, vol. 52, no. 7, pp. 777–788, Nov. 2015, doi: 10.1016/j.im.2015.02.003.
- [2] F. A. Shaqra, R. Duwairi, and M. Al-Ayyoub, "Recognizing emotion from speech based on age and gender using hierarchical models," *Procedia Computer Science*, vol. 151, pp. 37–44, 2019, doi: 10.1016/j.procs.2019.04.009.
- [3] J. Bhaskar, K. Sruthi, and P. Nedungadi, "Hybrid approach for emotion classification of audio conversation based on text and speech mining," *Procedia Computer Science*, vol. 46, pp. 635–643, 2015, doi: 10.1016/j.procs.2015.02.112.
- [4] M. T. P. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," *Procedia Computer Science*, vol. 108, pp. 1175–1184, 2017, doi: 10.1016/j.procs.2017.05.025.
- [5] D. Li, Y. Zhou, Z. Wang, and D. Gao, "Exploiting the potentialities of features for speech emotion recognition," *Information Sciences*, vol. 548, pp. 328–343, Feb. 2021, doi: 10.1016/j.ins.2020.09.047.
- [6] S. A. A. Thomas, and D. Mathew, "Study of MFCC and IHC feature extraction methods with probabilistic acoustic models for speaker biometric applications," *Procedia Computer Science*, vol. 143, pp. 267–276, 2018, doi: 10.1016/j.procs.2018.10.395.
- [7] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech-A review," in *Toward Robotic Socially Believable Behaving Systems - Volume 1: Modeling Emotions*, A. Esposito and L. C. Jain, Eds. Cham: Springer International Publishing, 2016, pp. 205–238.
- [8] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, Jan. 2020, doi: 10.1016/j.specom.2019.12.001.
- [9] J. Ancilin and A. Milton, "Improved speech emotion recognition with Mel frequency magnitude coefficient," *Applied Acoustics*, vol. 179, Aug. 2021, doi: 10.1016/j.apacoust.2021.108046.
- [10] H. Aouani and Y. Ben Ayed, "Speech emotion recognition with deep learning," *Procedia Computer Science*, vol. 176, pp. 251–260, 2020, doi: 10.1016/j.procs.2020.08.027.
- [11] H. Murugan, "Speech emotion recognition using CNN," *International Journal of Psychosocial Rehabilitation*, vol. 24, no. 8, pp. 2408–2416, 2020, doi: 10.37200/IJPR/V24I8/PR280260.
- [12] G. Sharma, K. Umopathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, 107020, Jan. 2020, doi: 10.1016/j.apacoust.2019.107020.
- [13] R. Y. Rumagit, G. Alexander, and I. F. Saputra, "Model comparison in speech emotion recognition for Indonesian language," *Procedia Computer Science*, vol. 179, pp. 789–797, 2021, doi: 10.1016/j.procs.2021.01.098.
- [14] A. Bin Abdul Quayyum, A. Arefeen, and C. Shahnaz, "Convolutional neural network (CNN) based speech-emotion recognition," in *2019 IEEE International Conference on Signal Processing, Information, Communication Systems (SPICSCON)*, 2019, pp. 122–125, doi: 10.1109/SPICSCON48833.2019.9065172.
- [15] M. G. de Pinto, M. Polignano, P. Lops, and G. Semeraro, "Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients," in *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, May 2020, pp. 1–5, doi: 10.1109/EAIS48028.2020.9122698.
- [16] J.-H. Yeh, T.-L. Pao, C.-Y. Lin, Y.-W. Tsai, and Y.-T. Chen, "Segment-based emotion recognition from continuous Mandarin Chinese speech," *Computers in Human Behavior*, vol. 27, no. 5, pp. 1545–1552, Sep. 2011, doi: 10.1016/j.chb.2010.10.027.
- [17] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Dec. 2016, pp. 1–4, doi: 10.1109/APSIPA.2016.7820699.
- [18] B. Zhang, C. Quan, and F. Ren, "Study on CNN in the recognition of emotion in audio and images," in *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, Jun. 2016, pp. 1–5, doi: 10.1109/ICIS.2016.7550778.
- [19] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv.1511.08458*, Nov. 2015.
- [20] R. Zhu, X. Tu, and J. X. Huang, "Deep learning on information retrieval and its applications," in *Deep Learning for Data Analytics*, H. Das, C. Pradhan, and N. Dey, Eds. Elsevier, 2020, pp. 125–153.
- [21] A. Pardamean and H. F. Pardede, "Tuned bidirectional encoder representations from transformers for fake news detection," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 22, no. 3, pp. 1667–1671, 2021, doi: 10.11591/ijeecs.v22.i3.pp1667-1671.
- [22] 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.
- [23] M. Gao, J. Dong, D. Zhou, X. Wei, and Q. Zhang, "Speech emotion recognition based on convolutional neural network and feature fusion," in *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, 2019, pp. 1145–1150, doi: 10.1109/ISKE47853.2019.9170369.




- [24] X. Ying, "An overview of overfitting and its solutions," *Journal of Physics: Conference Series*, vol. 1168, Feb. 2019, doi: 10.1088/1742-6596/1168/2/022022.
- [25] W. Dai and D. Berleant, "Benchmarking robustness of deep learning classifiers using two-factor perturbation," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 5085–5094, doi: 10.1109/BigData52589.2021.9671976.
- [26] 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, May 2018, doi: 10.1371/journal.pone.0196391.
- [27] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," *Scholars Portal Dataverse*, vol. 1, 2020, doi: 10.5683/SP2/E8H2MF.

## BIOGRAPHIES OF AUTHORS



**Fauziy Reggiswarashari**    is a fresh graduate with a bachelor's degree in informatics from Universitas Sebelas Maret who grows her interest in affective computing. Through reading and discussions, Fauziy found a bigger picture of affective computing implementation. By then, studying developments of affective computing challenges herself to learn more. She can be contacted at email: [xxfauziy@student.uns.ac.id](mailto:xxfauziy@student.uns.ac.id).



**Sari Widya Sihwi**    holds bachelor's and master's degree from the Faculty of Computer Science, Universitas Indonesia. Currently, she works as a lecturer at the Department of Informatics, Universitas Sebelas Maret, a public university in Indonesia. Her research interests are in several research areas in computer science, including affective computing. She can be contacted at email: [sariwidya@staff.uns.ac.id](mailto:sariwidya@staff.uns.ac.id).