Multi-Features Audio Extraction for Speech Emotion Recognition Based on Deep Learning

Jutono Gondohanindijo, Muljono*, Edi Noersasongko, Pujiono, De Rosal Moses Setiadi
Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia

***Abstract*—The increasing need for human interaction with computers makes the interaction process more advanced, one of which is by utilizing voice recognition. Developing a voice command system also needs to consider the user's emotional state because the users indirectly treat computers like humans in general. By knowing the type of a person's emotions, the computer can adjust the type of feedback that will be given so that the human-computer interaction (HCI) process will run more humanely. Based on the results of previous research, increasing the accuracy of recognizing the types of human emotions is still a challenge for researchers. This is because not all types of emotions can be expressed equally, especially differences in language and cultural accents. In this study, it is proposed to recognize speech-based emotion types using multi-feature extraction and deep learning. The dataset used is taken from the RAVDESS database. The dataset was then extracted using MFCC, Chroma, Mel-Spectrogram, Contrast, and Tonnetz. Furthermore, in this study, PCA (Principal Component Analysis) and Min-Max Normalization techniques will be applied to determine the impact resulting from the application of these techniques. The data obtained from the pre-processing stage is then used by the Deep Neural Network (DNN) model to identify the types of emotions such as calm, happy, sad, angry, neutral, fearful, surprised, and disgusted. The model testing process uses the confusion matrix technique to determine the performance of the proposed method. The test results for the DNN model obtained the accuracy value of 93.61%, a sensitivity of 73.80%, and a specificity of 96.34%. The use of multi-features in the proposed method can improve the performance of the model's accuracy in determining the type of emotion based on the RAVDESS dataset. In addition, using the PCA method also provides an increase in pattern correlation between features so that the classifier model can show performance improvements, especially accuracy, specificity, and sensitivity.**

Keywords*—Deep learning; multi-features extraction; RAVDESS; speech emotion recognition

I. INTRODUCTION

Speech is a form of information transfer commonly used in everyday life [1]. In everyday conversation, speech can provide a lot of information, not only words but also the emotions speakers convey. Knowing the level or type of emotion the other person is talking to is very important in building conversations in social life [2]. By understanding a person's emotional type, treatment and attitude towards that person will be adjusted to the current emotional state [3].

With the development of information technology and the increasing need for Human-Computer Interaction (HCI), the interaction process has become more advanced. One form of simple but effective advanced interaction is through speech [4][5]. The development of a voice command system needs to consider the user's emotional state, because when interacting with a computer, users tend to treat computers like humans in general [6][7]. Therefore, developing a sophisticated HCI system requires the availability of a speech database that represents emotions as a basis for developing an artificial intelligence system capable of imitating human emotions.

Speech Emotion Recognition (SER) is a field of research that focuses on recognizing types of human emotions which can then be processed further in the form of feedback to users. Several studies have carried out research related to SER, one of which was put forward by Chowdary and Hemanth [8], who utilized the Mel Frequency Cepstral Coefficients (MFCC) and Convolutional Neural Network (CNN) extraction methods to identify types of emotions using the RAVDESS database. This research produced 92%, 95%, and 69% accuracy, specificity, and sensitivity values, respectively. Another study conducted by Kumala and Zahra [9] proposed a study using a cross-corpus technique to identify the types of emotions in Indonesian conversation. This study's results indicate that using a combination of two SER databases and feature extraction of MFCC and Teager Energy can provide an accuracy of 85.42%. Based on the achievement of the performance parameters of several studies above, it can be said that increasing the accuracy of recognition of types of human emotions is still a challenge for researchers. This happens because not all types of emotions can be expressed equally [10].

From the research above, it can be seen that the results of speech-based emotion recognition depend heavily on the database used, the number of balanced classes, the feature extraction process, and the machine learning method used [11]. The use of multiple features is one aspect of improving the performance of the classifier model in identifying types of human emotions, as stated in the research proposed by Iqbal et al [12], where this research utilizes the features of frequency, Pitch, Amplitude, and Formant which are combined with ANN models based on Bayesian Regularized (BRANN) to recognize the type of emotion using the Berlin Database of Emotional Speech (Berlin EmoDB) dataset. The evaluation results of this study obtained an accuracy value of 95%. These results indicate that the use of several features in recognizing the types of human emotions can have an impact, especially in increasing the value of performance parameters in the classifier model.

Therefore, this study proposes using multi-feature extraction and deep learning methods by utilizing the Deep Neural Network (DNN) architecture to recognize types of emotions based on speech. Deep learning is used because of its ability to process large data and provide high-accuracy values [13]. Furthermore, the multi-features used in this study consist of Mel Frequency Cepstral Coefficients (MFCC), Chroma, Mel-Spectrogram, Tonnetz, and Contrast. These five features are related to the high and low frequency of speech associated with emotional expression [14]. In addition, PCA (Principal Component Analysis) and Min-Max Normalization techniques will be applied to determine the impact resulting from the application of these techniques. The evaluation results were then analyzed using the Confusion Matrix table to determine the accuracy, specificity, and sensitivity values.

This speech emotion recognition research contributes to:

1) Determine the features that affect the recognition of eight (8) classes of human speech emotions

2) Determine the optimal number of features for feature selection using PCA and feature normalization using MinMax

3) Improve performance parameters of accuracy, sensitivity and specificity.

The following section will explain related research, especially research on SER. Then, Section III will explain the methodology used in this study. Then in the next section, we will present the results of our experiment along with an analysis of these results, and Section V is this study's conclusion.

II. Related Research

SER, or Speech Emotion Recognition, is a method of recognizing types of emotions through speech by utilizing several data processing techniques and machine learning. Based on the results of a literature study conducted by Singh and Goel [11], it shows that SER is a research area that has high interest, especially in real-world applications. From the results of this study, it was found that the development of SER has several factors that influence the results and performance of SER, namely the availability of datasets used in model development, the feature extraction process that is relevant to the type of emotion, and the type of classifier used in model training.

Several studies have carried out the development of models related to SER, one of which was suggested by research conducted by Chowdary and Hemanth [8]. This study proposes the development of SER by utilizing the RAVDESS and Convolutional Neural Network (CNN). The research phase starts from the MFCC feature extraction stage from the RAVDESS dataset. Then the feature extraction results were validated into 1642 data as training data and 810 data as test data. The training data is used as a reference for training the Conv1D-based CNN model. Next, the model is evaluated using test data. The evaluation results showed a model accuracy of 92%, sensitivity of 69%, and specificity of 95%.

Furthermore, Iqbal et al. [12] proposed speech emotion recognition based on Artificial Neural Networks (ANN). The Berlin Database of Emotional Speech (Berlin EmoDB) was used in this study as a speech-based emotion recognition dataset collection. Several feature extractions are used, including frequency, pitch, amplitude, and formant. While the classifier model used is ANN based on Bayesian Regularized (BRANN). The model evaluation results obtained an accuracy performance of 95%. In addition, several studies also utilize multi-feature techniques such as MFCC [15], Cross Zero Rate, Root Mean Square (RMS) [16], Chroma, Mel-Spectrogram, Contrast, and Tonnetz [17][14]. The use of multiple features is one aspect of increasing the performance of the classifier model in identifying types of emotions [11].

The use of a different approach was also proposed by Kumala and Zahra [9]. In this study, it is proposed to use the Cross-Corpus technique to recognize speech emotions in Indonesian. This study utilizes several database sources, namely the Berlin Database of Emotional Speech (Berlin EmoDB), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and Surrey Audio-Visual Expressed Emotion (SAVEE) so that there are three corpora, consisting of one corpus in German, and two corpora in English. The feature extraction process in this study uses the MFCC and Teager Energy methods. Using the Support Vector Machine (SVM) classifier, this study increased accuracy performance by 4.16% on MFCC and 2.09% on the Teager-MFCC combination. In addition, the three corpora used are in good agreement with Indonesian in terms of emotion recognition.

Using the multi-acoustic feature on the SER is one of the efforts to improve the performance of human emotion recognition. This study will use features such as MFCC, Chroma, Contrast, Mel-Spectrogram, and Tonnetz as emotion recognition features. Then Principal Component Analysis (PCA) and Min-Max Normalization techniques are implemented to see the impact of these processes on the performance of emotion recognition. Deep Neural Network (DNN) is used in this study as a classifier because of its ability to process large data and provide high accuracy values [13]. The model that has been trained is then evaluated using the confusion matrix to determine the performance of the model, especially in terms of accuracy, specificity and sensitivity values.

III. Methodology

This study proposes to identify emotions based on RAVDESS voice data using multi-feature extraction and Deep Neural Network (DNN). An overview of this research can be seen in Fig. 1.

In Fig. 1 it can be seen that the RAVDESS dataset will extract its audio features using several types of audio extraction methods such as Mel Frequency Cepstral Coefficients (MFCC), Chroma, Mel-Spectrogram, Tonnetz, and Contrast. Then, PCA (Principal Component Analysis) and Min-Max Normalization techniques are applied to the extracted features to determine the impact of transformation and data reduction on the resulting model.

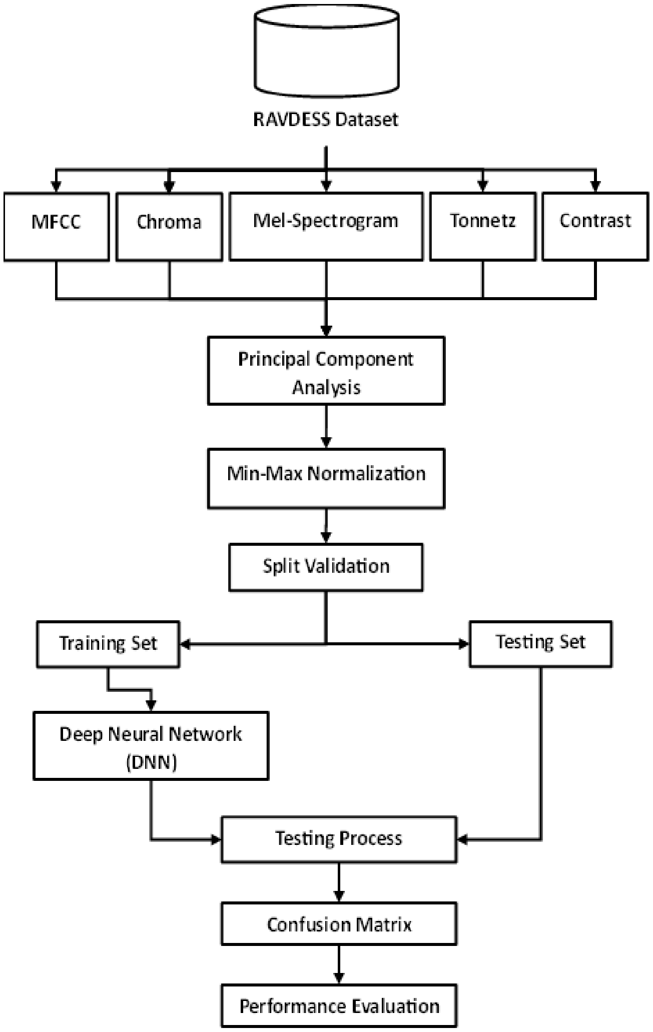


Fig. 1. Concept of proposed study.

After experiencing the pre-processing stage, the dataset can be grouped into training and testing sets using split validation. The Deep Neural Network (DNN) will use the training set as the model training reference data. After the training results model is obtained, the next step is to test using the testing set as the test data. From the testing process, the results of the emotion type prediction will be obtained by the trained DNN model, where these results will be transformed into a Confusion Matrix table as a reference table for determining the performance parameters of the proposed model. The parameters used to determine the performance of the proposed model are accuracy, specificity, and sensitivity.

A. *Dataset*

In this study, the Ryerson Audio-Visual Database of Emotional Speech and Song, or the RAVDESS dataset, was used [18]. The data set is a multi-modal database consisting of separate sound and video recordings showing certain types of emotions. The database is gender balanced, consisting of 24 professional actors, vocalizing lexically-matched statements in a neutral North American accent. Each actor recites a sentence or song in North American English lasting 3 to 4 seconds in a standardized recording scenario.

In this study, the focus will be on datasets in audio format. RAVDESS contains 2452 audio data divided into 1440 audio data for speech and 1012 audio data for songs. The data is also divided into 8 (eight) types of emotions: calm, happy, sad, angry, neutral, fearful, surprise, and disgust.

B. *Feature Extraction*

1) Mel Frequency Cepstral Coefficients (MFCC) is a feature extraction type commonly used in audio files [19]. MFCC is generally suggested to be used as an identifier for monosyllables in audio without identifying the speaker [20]. The MFCC feature extraction process [8] in audio begins with the Pre-emphasis stage, namely amplifying the audio signal at high frequencies. Followed by the framing and windowing stages, where framing stage aims to divide the length of the audio into several time intervals between 20 ms to 30 ms while the windowing technique is used to limit the occurrence of disturbances at the beginning and end of the audio, the next stage is the implementation of the Fast Fourier Transform, Mel Filter Bank, and Discrete Cosine Transform as a process of transforming the windowing results into MFCC. MFCC (Mel-Frequency Cepstral Coefficients) is a feature used in speech emotion recognition which has the advantage of representing the acoustic properties of the human voice. The MFCC uses the mel scale, which is similar to the human auditory perception of frequency. MFCC features are generated by taking the logarithm of the power spectrum and converting it to cepstrum, thereby helping to reduce feature dimensionality and processing complexity. MFCC can represent temporal information in speech signals through short-duration frame splitting techniques to capture variations in speech signals associated with temporal emotional changes.

2) Chroma is a feature extraction focusing on music-oriented audio tones [21]. This feature can provide a distribution of tonal variations in audio in the form of a simple feature. The Chroma feature's result is a chromagram built based on 12 (twelve) tone levels [22]. The use of chroma is expected to recognize the high and low pitch of the actor's speech in audio, where the tone of the speech can indicate a certain type of emotion.

3) Mel-Spectrogram is an audio feature extraction that was built to overcome the problem of limited human hearing ability in distinguishing high-frequency values [22]. The use of the Mel-Spectrogram in this study is to extract information on differences in frequency values, particularly in identifying the types of emotions expressed by actors.

4) Tonnetz is a feature extraction derived from Chroma that also focuses on audio harmony and tone classes [23].

5) Contrast is a feature extraction in audio that is useful for estimating the average sound energy based on each sub-band's peak and valley spectral values [24].

C. *Principal Component Analysis (PCA)*

PCA or Principal Component Analysis, is a statistical method that is widely used, especially in data processing such as dimension reduction, data compression, and feature extraction [25]. PCA is conceptually able to identify new variables based on the main components, where these values are linearly the result of the combination of the original features used [26]. Simply put, PCA will project a new feature or variable whose representation is the same as the original feature where the number of components can be adjusted. In this study, PCA will be tested as dimension reduction to reduce the number of extracted features and increase the representation of feature values.

D. Min-Max Normalization

Normalization is the process of equalizing values among features with significant differences in value so that the weights and the effects on each feature are the same when used as a reference for classifier model training [27]. In this study, the Min-Max Normalization method will be applied where the value of each feature will be distributed over a range of values between 0 and 1. The application of this method will provide an overview of the impact of using normalization in the formation of classifier models. This method works by first determining the maximum (x_{max}) and minimum (x_{min}) values of each variable or feature. Then each original data (x_{old}) is operated with the previously obtained value to produce a new value (x_{new}) using the following equation [8]:

$$x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}} \tag{1}$$

E. Deep Neural Network (DNN)

Deep Neural Network or DNN is one of the Deep Learning (DL) methods built on the basis of Neural Networks. DNN is the improved version of the conventional Neural Network method by adding some depth such as additional hidden layers at the input and output layers [28]. This method is generally used to predict or classify data according to class. In this study, the DNN structure used consisted of 1 (one) dense layer as input and 1 (one) dense layer as output with each activation, namely 'ReLu' and 'Softmax'. Then there are 3 (three) hidden solid layers. A detailed description of the proposed DNN structure in this study can be seen in Table I:

TABLE I. PROPOSED DNN STRUCTURE

Component	Layer Architecture	Activation Function
Input Layer	Dense Layer	RELU
1st Hidden Layer	Dense Layer	RELU
Dropout	DROPOUT	-
2nd Hidden Layer	Dense Layer	RELU
Dropout	DROPOUT	-
3rd Hidden Layer	Dense Layer	RELU
Dropout	DROPOUT	-
Output Layer	Dense Layer	SOFTMAX

F. Performance Evaluation

The DNN model testing process results will be converted into a Confusion Matrix table as a reference for calculating model performance parameters. The performance parameters used are accuracy, specificity, and sensitivity. Determining the performance parameter values of the proposed model can use the following equation [8] :

$$Accuracy = \frac{TN+TP}{FP+FN+TN+TP} \tag{2}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{3}$$

$$Specificity = \frac{TN}{TN+FP} \tag{4}$$

In the equation above, TP (True Positive) is the number of test data correctly predicted as a positive class, TN (True Negative) is the amount of test data correctly predicted as a negative class while FP (False Positive) is the amount of test data with a negative class which is predicted as the positive class and FN (False Negative) is the number of test data with the positive class which is predicted to be the negative class. These four values can be generated from the Confusion Matrix table.

IV. EXPERIMENT RESULT AND DISCUSSION

Following are the steps in implementing the use of the RAVDESS dataset for speech emotion recognition:

1) Dataset exploration and understanding: understand the structure, metadata, and emotional information provided and examine the number of sound recordings, actors involved.

2) Feature extraction: uses the combined extraction technique and determines the feature extraction parameters.

3) Feature Selection: using the PCA technique

4) Data Normalization: using the MinMax Technique

5) Dataset division: divide the RAVDESS dataset into training, validation, and testing subsets.

6) Model training and evaluation: train a speech emotion recognition model using training subsets and validation splits.

7) Testing and final validation: using a subset of testing to test the model that has been trained and calculating the confusion of testing metrics to get accuracy, sensitivity, specificity.

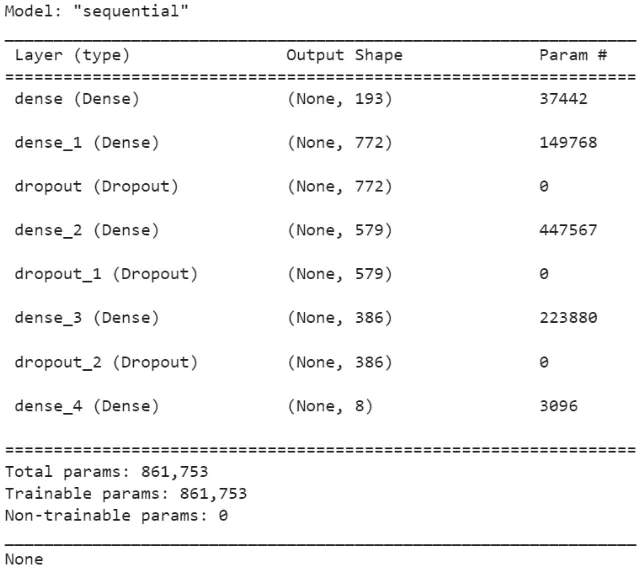
8) Analysis of results and evaluation: analyze the results of speech emotion recognition obtained from the model, including performance in each emotion category.

In this study, the experimental phase was carried out using RAVDESS audio dataset which consisted of 1440 spoken audio data and 1012 emotional song data in *.wav format. Furthermore, the data were extracted using several feature extraction techniques, consisting of MFCC, Mel-Spectrogram, Chroma, Contrast, and Tonnetz. From the extraction results that have been carried out, a total of 193 features were obtained consisting of 40 MFCC features, and 12 Chroma features, whereas Mel-Spectrogram, Contrast, and Tonnetz produced a total of 128 features, 7 features, and 6 features, respectively.

Furthermore, the features obtained are processed using the PCA technique with the total components used are the multiple of 10% of the total features. These results are then normalized using the Min-Max Normalization method to limit the data range that is too large, so that the same range of data for each feature can be achieved.

In the next stage, the normalization results using Min-Max Normalization were validated using separate validation with a ratio of 33% for 810 data as a test set and the remaining 1642 data as a training set. Next, the initialization of the sequential model is carried out by utilizing dense layers to form a Deep Neural Network (DNN) model.

The Deep Neural Network (DNN) model used consists of 5 (five) dense layers. Then the parameter specifications for each layer used consist of the first to fourth dense layers using the 'ReLu' activation function, while the fifth (last) dense layer uses the 'Softmax' activation function which acts as an inference in determining emotion classes. The proposed DNN model uses the 'Adam' optimizer parameter or the adaptive estimates of lower-order moments [29]. Then, the value of 0.1 is used for the dropout rate parameter, which means that the ratio of possible elimination nodes in the DNN is 10% at each embedded dropout step. The model of this sequential DNN uses dropout and parameters, which can be seen in Fig. 2.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 193)               37442

 dense_1 (Dense)             (None, 772)               149768

 dropout (Dropout)           (None, 772)               0

 dense_2 (Dense)             (None, 579)               447567

 dropout_1 (Dropout)         (None, 579)               0

 dense_3 (Dense)             (None, 386)               223880

 dropout_2 (Dropout)         (None, 386)               0

 dense_4 (Dense)             (None, 8)                 3096

=================================================================
Total params: 861,753
Trainable params: 861,753
Non-trainable params: 0
_________________________________________________________________
None
```

Fig. 2. Specification of deep neural network (DNN) using 100% of features.

DNN model training using 1642 data distributed into 8 (eight) classes, namely calm, happy, sad, angry, neutral, afraid, surprised, and fed up has been carried out. The training process is repeated for 200 epochs. The model produced in the training process was tested using test data with 810 data. The experimental scheme in this study was carried out using several combinations of data pre-processing methods, especially PCA and Min-Max Normalization. The variations of the experimental schemes used consist of experiments with original features, experiments with applying the normalization method, experiments with applying the PCA technique, and experiments with a combination of PCA and Normalization techniques. In experiments that apply the PCA technique, the number of components used is 100% to 10% of the total features used, where the decrease in the number of components used is 10% for each test. Then, the experimental results for each scheme are transformed into a confusion matrix and evaluated using the performance parameters as shown in Table II.

Table II shows the test results for each experimental variation of the proposed method. The initial scheme that uses all original features can produce a Sensitivity of 69.53%, Specificity of 95.93%, and Accuracy of 92.93%. This table also shows that the use of the Min-Max Normalization and PCA methods can have an impact on the performance value of the classifier model.

TABLE II. THE RESULT OF VARIATION EXPERIMENTAL OF THE PROPOSED METHOD

	Sensitivity (%)	Specificity (%)	Accuracy (%)
Original Features	69,53	95,93	92,93
Normalization	67,71	95,55	92,28
PCA	**73,80**	**96,34**	**93,61**
PCA + Normalization	73,00	96,33	93,61

The use of the Min-Max Normalization method in this study impacts decreasing performance values, although the decrease is not too significant. This can happen because Min-Max Normalization only projects the original feature values to be valued from 0 to 1, so there is a potential for important values to be omitted, which results in bias during pattern analysis in the model-building process.

The application of the PCA method can impact increasing the performance value of schemes with original features. In fact, applying the PCA method provided the best overall experimental performance with an accuracy value of 93.61%, a Sensitivity of 73.80%, and a Specificity of 96.34%. These results were obtained using 100% components or 193 components of PCA. The achievement of this value can occur because at the PCA stage, there is a Data Scaling process, which is similar to the transformation of feature values in the Normalization method. Furthermore, a statistical calculation process is carried out to form component values close to the original feature values, so that the distribution of feature values becomes the same and PCA can also increase the correlation of each component.

Then in the experiment that combined PCA and Min-Max Normalization, the highest results were obtained with sensitivity of 73%, specificity of 96.33% and accuracy of 93.61%. These results have similarities with the results achieved by experiments using PCA only. This can be indicated that the addition of normalization techniques to combined experiments between PCA and Normalization tends to have an impact in the form of decreasing the achievement of model performance values.

Furthermore, Table III displays the result of the PCA technique where a trial iteration is carried out with a 10% reduction in the number of components. It shows a change in the model performance value. As shown in Table III, the drastic reduction of each performance parameter starts from 50% of the components used or half of the total number of features.

This can happen because changes in the number of features or components affect the pattern analysis results from the classifier. The lower the number of PCA components used, the lower the model's accuracy, sensitivity and specificity performance. In the graphic, it can be seen in Fig. 3, 4, and 5.

TABLE III. THE IMPACT OF THE REDUCTION OF PCA FEATURES ON MODEL PERFORMANCE

Component Percentage (%)	Number of Features	PCA			PCA + Normalization		
		Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)
100%	193	**73.80**	**96.34**	**93.61**	73.00	96.33	93.61
90%	174	73.24	96.24	93.49	72.07	96.12	93.27
80%	155	72.44	96.10	93.21	73.12	96.24	93.46
70%	136	70.64	96.04	93.15	72.34	96.19	93.40
60%	116	71.32	95.98	93.02	70.45	95.91	92.90
50%	97	69.36	95.75	92.59	67.52	95.58	92.31
40%	78	67.24	95.47	92.16	67.71	95.51	92.22
30%	58	66.34	95.30	91.85	61.62	94.69	90.77
20%	39	58.21	94.30	90.09	58.08	94.20	89.97
10%	20	52.42	93.50	88.67	49.39	92.97	87.72

Fig. 3, 4, and 5 respectively explain the decrease in Accuracy (Fig. 3), Sensitivity (Fig. 4) and Specificity (Fig. 5) performance when using PCA or combined PCA + Normalization at each stage of the performance test with a feature reduction of 10% for each stage.

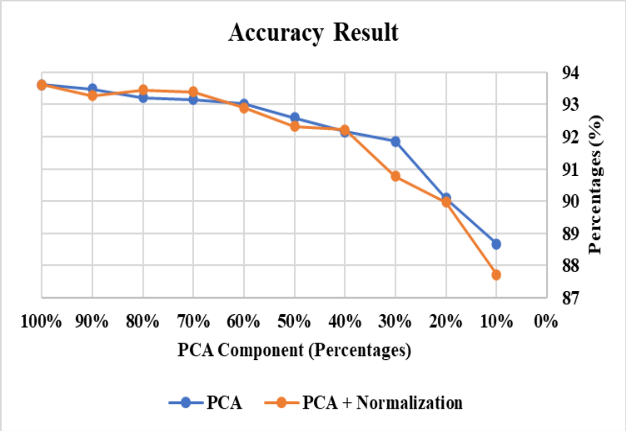


Fig. 3. Accuracy performance.

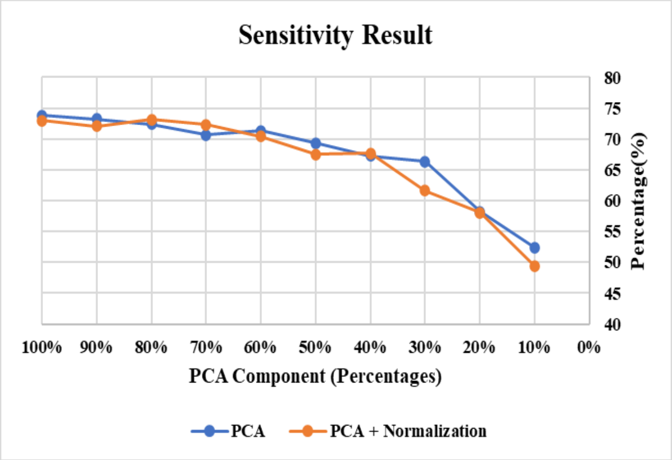


Fig. 4. Sensitivity performance.

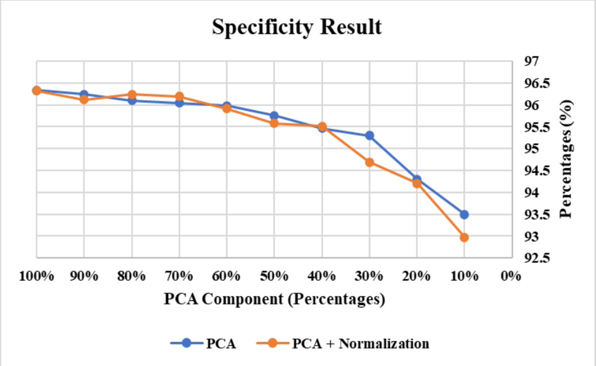


Fig. 5. Specificity performance.

Overall, it was found that high-performance results were obtained with experiments implementing PCA only. Table IV shows a detailed description of the achievement of performance parameter values for each type of emotion in the experiment.

In Table IV, it can be seen that several types of emotions have an accuracy value above 94%, including happy, sad, fearful, and surprised. While the types of disgust and anger emotions obtained an accuracy value of 93.70% and 92.96%, respectively, followed by neutral and calm emotions with an accuracy value of 91.60%. The accuracy value is an indicator that shows how the model performs in predicting a data class correctly. The evaluation results show several differences in the accuracy value in each class. This can occur due to differences in the amount of data distribution according to each class [30].

Furthermore, the specificity parameter is a parameter that indicates the classifier's ability to predict a negative class among all data with a negative class. From the results of the specificity test, the proposed model is capable of producing performance above 95% with an average performance of 96.34%, so it can be said that the proposed model is capable of producing high specificity values.

TABLE IV. THE PERFORMANCE RESULT OF EACH EMOTION CLASS

Class	TP	TN	FP	FN	Sensitivity	Specificity	Accuracy
Neutral	89	653	26	42	67.94%	96.17%	91.60%
Calm	91	651	34	34	72.80%	95.04%	91.60%
Happy	48	718	24	20	70.59%	96.77%	94.57%
Sad	45	722	29	14	76.27%	96.14%	94.69%
Angry	88	665	27	30	74.58%	96.10%	92.96%
Fearful	110	657	21	22	83.33%	96.90%	94.69%
Disgust	92	667	27	24	79.31%	96.11%	93.70%
Surprised	40	730	19	21	65.57%	97.46%	95.06%
Average					**73.80%**	**96.34%**	**93.61%**

Then the sensitivity value is a parameter that shows the model's ability to predict the positive class correctly among all data that is in the positive category. The test results show that the sensitivity value for the type of fearful emotion has a value above 83%, indicating that the proposed model can identify test data with the type of fearful emotion well. Meanwhile, other types of emotions such as neutral, calm, happy, sad, angry, and surprised, can produce a sensitivity value of less or a little bit more than 70%, with an average value for all emotion classes of 73.80%. These results indicate that the RAVDESS dataset has a fairly high level of bias between classes. This can occur because the expression of several types of emotions tends to differ between actors [10].

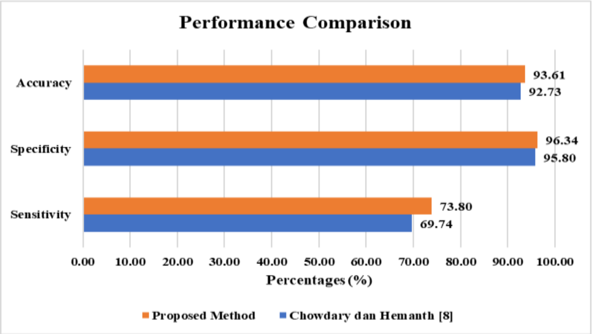


Fig. 6. Comparison of performance results with previous study.

Overall, the performance evaluation results of the proposed DNN model were able to produce the average values of accuracy, sensitivity and specificity of 93.61%, 73.80% and 96.34%, respectively, it can be seen in Fig. 6. These results were able to outperform the results of previous studies put forward by Chowdary and Hemanth [8], where this study also used RAVDESS dataset with MFCC feature extraction and CNN classifier. Comparison of performance results can be seen in Table V.

Based on Table V, the table compares the accuracy value of the proposed method with several previous studies. The comparison of the accuracy values used is the result using the RAVDESS dataset only. From Fig. 7, it can be seen that the proposed method is able to outperform the performance of previous studies. The use of multi-features in the proposed method can improve the performance of the model's accuracy in determining the type of emotion based on the RAVDESS dataset. In addition, the use of the PCA method also provides an increase in pattern correlation between features so that the classifier model can show performance improvements, especially accuracy, specificity, and sensitivity values.

TABLE V. THE PERFORMANCE RESULT OF EACH EMOTION CLASS

No.	Work	Dataset	Feature	Classifier	Accuracy Result
1	Chowdary and Hemanth [8]	RAVDESS	Mel Frequency Cepstral Coefficients (MFCC)	CNN	92%
2	Jothimani and Premalatha [16]	RAVDESS, CREMA, SAVEE, and TESS	Mel Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), and Root Mean Square (RMS)	CNN+LSTM	92.60%
3	Alnuaim et al [31]	RAVDESS	Mel Frequency Cepstral Coefficients (MFCC), Short-time Fourier transform and Mel Spectrogram	MLP classifier	81%
4	Patnaik [32]	RAVDESS and TESS	Complex Mel Frequency Cepstral Coefficients (c-MFCC)	deep sequential LSTM model	91.60%
5	Proposed Method	RAVDESS	Mel Frequency Cepstral Coefficients (MFCC), Chroma, Mel Spectogram, Constrast, Tonnetz	Principal Component Analysis (PCA) Deep Neural Network (DNN)	93.61%

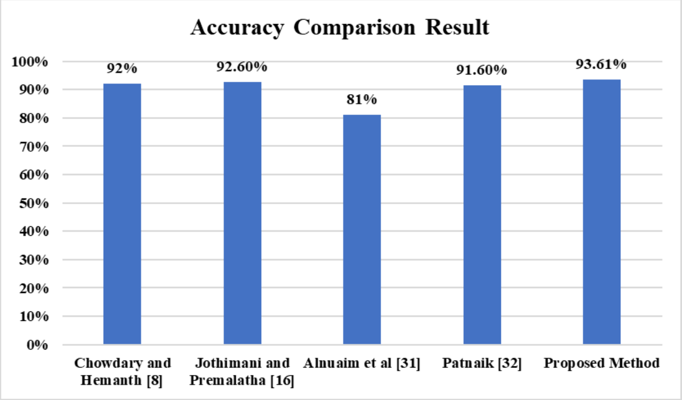


Fig. 7. Accuracy comparison between proposed method with previous studies.

V. Conclusion

Speech Emotion Recognition (SER) based on multi-feature extraction and Deep Neural Network (DNN) has been carried out. A total of 2452 audio data in .wav format taken from the RAVDESS database were used in this study. The data is extracted to produce several features, including Mel Frequency Cepstral Coefficients (MFCC), Chroma, Mel-Spectrogram, Contrast, and Tonnetz. From the extraction process, 193 main features were obtained. This study examines the impact of applying Principal Component Analysis (PCA) and Min-Max Normalization to the performance of the classifier model used. The DNN model is used in this study to determine emotions such as calm, happy, sad, angry, neutral, fearful, surprised, and disgusted. The test results for the DNN model with 200 epochs were able to obtain the accuracy of 93.61%, sensitivity of 73.80%, and specificity of 96.34%. The use of multiple features in the proposed method can improve the model's accuracy in determining the type of emotion based on the RAVDESS dataset. In addition, using the PCA method also provides an increase in pattern correlation among features so that the classifier model can show performance improvements, especially accuracy, specificity, and sensitivity. Moreover, the scheme that uses the PCA technique in which experimental iterations are carried out by reducing the number of components by 10% shows a change in the value of the model's performance, especially when the model uses features less than 50% of all components. The lower the number of PCA components used, the lower the performance of the model. The implementation of multiple features in this study can open opportunities for using other features related to certain types of emotions. Furthermore, the use of other dataset and classifiers can also provide a new approach in the development of this research in the future.

References

[1] T. Puri, M. Soni, G. Dhiman, O. Ibrahim Khalaf, M. alazzam, and I. Raza Khan, “Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network,” J. Healthc. Eng., vol. 2022, no. ii, 2022, doi: 10.1155/2022/8472947.

[2] N. Ahmed, Z. Al Aghbari, and S. Girija, “A systematic survey on multimodal emotion recognition using learning algorithms,” Intell. Syst. with Appl., vol. 17, no. January, p. 200171, 2023, doi: 10.1016/j.iswa.2022.200171.

[3] M. Egger, M. Ley, and S. Hanke, “Emotion Recognition from Physiological Signal Analysis: A Review,” Electron. Notes Theor. Comput. Sci., vol. 343, pp. 35–55, 2019, doi: 10.1016/j.entcs.2019.04.009.

[4] W. Alsabhan, “Human–Computer Interaction with a Real-Time Speech Emotion Recognition with Ensembling Techniques 1D Convolution Neural Network and Attention,” Sensors, vol. 23, no. 3, p. 1386, Jan. 2023, doi: 10.3390/s23031386.

[5] Muljono, A. Q. Syadida, D. R. I. M. Setiadi, and A. Setyono, “Sphinx4 for Indonesian continuous speech recognition system,” in Proceedings - 2017 International Seminar on Application for Technology of Information and Communication: Empowering Technology for a Better Human Life, iSemantic 2017, 2017, pp. 264–267. doi: 10.1109/ISEMANTIC.2017.8251881.

[6] R. L. Soash, “Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places,” Collect. Manag., vol. 24, no. 3–4, pp. 310–311, 1999, doi: 10.1300/j105v24n03_14.

[7] C. Breazeal, “Emotion and sociable humanoid robots,” Int. J. Hum. Comput. Stud., vol. 59, no. 1–2, pp. 119–155, 2003, doi: 10.1016/S1071-5819(03)00018-1.

[8] M. Kalpana Chowdary and D. Jude Hemanth, “Deep Learning Approach for Speech Emotion Recognition,” in Data Analytics and Management, 2021, pp. 367–376. doi: 10.1007/978-981-15-8335-3_29.

[9] O. U. Kumala and A. Zahra, “Indonesian Speech Emotion Recognition using Cross-Corpus Method with the Combination of MFCC and Teager Energy Features,” Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 4, pp. 163–168, 2021, doi: 10.14569/IJACSA.2021.0120422.

[10] A. Chowanda and Y. Muliono, “Emotions Classification from Speech with Deep Learning,” Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 4, pp. 777–781, 2022, doi: 10.14569/IJACSA.2022.0130490.

[11] Y. B. Singh and S. Goel, “A systematic literature review of speech emotion recognition approaches,” Neurocomputing, vol. 492, pp. 245–263, Jul. 2022, doi: 10.1016/j.neucom.2022.04.028.

[12] M. Iqbal, S. A. Raza, M. Abid, F. Majeed, and A. A. Hussain, “Artificial Neural Network based Emotion Classification and Recognition from Speech,” Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 12, pp. 434–444, 2020, doi: 10.14569/IJACSA.2020.0111253.

[13] T. J. Saleem and M. A. Chishti, “Deep learning for the internet of things: Potential benefits and use-cases,” Digit. Commun. Networks, vol. 7, no. 4, pp. 526–542, 2021, doi: 10.1016/j.dcan.2020.12.002.

[14] B. Pragati, C. Kolli, D. Jain, A. V. Sunethra, and N. Nagarathna, “Evaluation of Customer Care Executives Using Speech Emotion Recognition,” in Machine Learning, Image Processing, Network Security and Data Sciences, 2023, pp. 187–198. doi: 10.1007/978-981-19-5868-7_14.

[15] K. Nugroho, E. Noersasongko, Purwanto, Muljono, and H. A. Santoso, “Javanese Gender Speech Recognition Using Deep Learning and Singular Value Decomposition,” in Proceedings - 2019 International Seminar on Application for Technology of Information and Communication: Industry 4.0: Retrospect, Prospect, and Challenges, iSemantic 2019, 2019, pp. 251–254. doi: 10.1109/ISEMANTIC.2019.8884267.

[16] S. Jothimani and K. Premalatha, “MFF-SAug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network,” Chaos, Solitons & Fractals, vol. 162, p. 112512, Sep. 2022, doi: 10.1016/j.chaos.2022.112512.

[17] S. Patra, S. Datta, and M. Roy, “Analysis on Speech-Emotion Recognition with Effective Feature Combination,” in 2022 OITS International Conference on Information Technology (OCIT), IEEE, Dec. 2022, pp. 1–5. doi: 10.1109/OCIT56763.2022.00018.

[18] 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.

[19] R. M. Hanifa, K. Isa, and M. Mohamad, “Comparative Analysis on Different Cepstral Features for Speaker Identification Recognition,”

2020 IEEE Student Conf. Res. Dev. SCOReD 2020, no. September, pp. 487–492, 2020, doi: 10.1109/SCOReD50371.2020.9250938.

[20] S. Ajibola Alim and N. Khair Alang Rashid, “Some Commonly Used Speech Feature Extraction Algorithms,” in From Natural to Artificial Intelligence - Algorithms and Applications, IntechOpen, 2018. doi: 10.5772/intechopen.80419.

[21] J. V. T. Abraham, A. N. Khan, and A. Shahina, “A deep learning approach for robust speaker identification using chroma energy normalized statistics and mel frequency cepstral coefficients,” Int. J. Speech Technol., no. 0123456789, 2021, doi: 10.1007/s10772-021-09888-y.

[22] U. Garg, S. Agarwal, S. Gupta, R. Dutt, and D. Singh, “Prediction of Emotions from the Audio Speech Signals using MFCC, MEL and Chroma,” Proc. - 2020 12th Int. Conf. Comput. Intell. Commun. Networks, CICN 2020, pp. 87–91, 2020, doi: 10.1109/CICN49253.2020.9242635.

[23] S. Sen, A. Dutta, and N. Dey, “Speech Processing and Recognition System,” in Audio Processing and Speech Recognition, 2019, pp. 13–43. doi: 10.1007/978-981-13-6098-5_2.

[24] S. Bhattacharya, S. Borah, B. K. Mishra, and A. Mondal, “Emotion detection from multilingual audio using deep analysis,” Multimed. Tools Appl., vol. 81, no. 28, pp. 41309–41338, 2022, doi: 10.1007/s11042-022-12411-3.

[25] T. Kurita, “Principal component analysis (PCA),” in Computer Vision: A Reference Guide, Springer, 2019, pp. 1–4. doi: 10.48550/arXiv.1503.06462.

[26] M. Ringnér, “What is principal component analysis?,” Nat. Biotechnol., vol. 26, no. 3, pp. 303–304, 2008.

[27] S. G. K. Patro and K. K. Sahu, “Normalization: A Preprocessing Stage,” IARJSET, pp. 20–22, Mar. 2015, doi: 10.17148/IARJSET.2015.2305.

[28] J.-T. Chien, “Deep Neural Network,” in Source Separation and Machine Learning, Elsevier, 2019, pp. 259–320. doi: 10.1016/B978-0-12-804566-4.00019-X.

[29] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” Dec. 2014.

[30] N. Dogan and Z. Tanrikulu, “A comparative analysis of classification algorithms in data mining for accuracy, speed and robustness,” Inf. Technol. Manag., vol. 14, no. 2, pp. 105–124, 2013, doi: 10.1007/s10799-012-0135-8.

[31] A. A. Alnuaim et al., “Human-Computer Interaction for Recognizing Speech Emotions Using Multilayer Perceptron Classifier,” J. Healthc. Eng., vol. 2022, 2022, doi: 10.1155/2022/6005446.

[32] S. Patnaik, “Speech emotion recognition by using complex MFCC and deep sequential model,” Multimed. Tools Appl., vol. 82, no. 8, pp. 11897–11922, 2023, doi: 10.1007/s11042-022-13725-y.