ISSN No:-2456-2165

Gender Detection by Voice Using Deep Learning

Mia Mutiany
Department of Informatics
WidyatamaUniversity
Bandung, Indonesia

Iwa Ovyawan Herlistiono
Department of Informatics
Widyatama University
Bandung, Indonesia

Abstract:- The recognition of gender voices as an important part of answering certain voices. To distinguish gender from sound signals, sound techniques have defined the gender-relevant features (male or female) of these sound signals. In this study, we used various models to improve accuracy, one of which was by using deep learning with the voice gender DNN method. This noise reduction uses the extraction feature of the Mel Frequency Cepstral Coefficient (MFCC), then the sound classification uses SVM. By using a separation ratio of 80% for training data and 20% for testing data. The results showed that using DNN for voice recognition was better and pairing with the SVM algorithm obtained an accurate result of 0.97%.

Keywords:- *Voice Recognition, Deep Neural Network, Deep Learning, MFCC, SVM.*

I. INTRODUCTION

Voice recognition is one of the important researches that are currently widely used from a variety of applications, such as security systems, authentication, and so on. Voice recognition must have high performance, to be able to improve speech recognition performance, one of which is by adding a gender classification procedure. With this gender classification, the problem space in speech recognition can be limited only based on predetermined gender[1].

Voice data is divided into training data and testing data by classifying gender into two categories, namely male and female. Male and female voices have their characteristics due to different resonances in the throat[2]. By processing sound signals, these characteristics will be obtained in a form that can be recognized by a computer. With thesecharacteristics, the computer can identify gender through sound signals. Therefore, we need a learning algorithm that can help humans detect sounds based on gender. The deep learning algorithm can support this detection because it can predict more accurately and quickly for speech recognition.

Deep Learning has excellent models for detections such as image recognition, emotional recognition, and speech recognition. Deep learning which is commonly used for speech recognition is a deep neural network (DNN). Therefore, deep learning with the DNN model will be used to detect speech recognition. The deep learning algorithm used for this study uses several existing feature extractions.

MFCC chose feature extraction because it is a fairly good feature extraction method for noise reduction which requires a fast, easy, and complete processing time. Meanwhile, the classification of votes is based on gender using a machine support vector (SVM). [3]

II. RELATED WORK

Several studies discussdetecting voices based on gender. Previous research conducted by Martin and Joensuu who developed speech recognition using the detected GMM and FFT features gave the best results for the classification level[4][5].

S.LYuan[6] developed a voice recognition system and detected gender based on voice using Deep Learning with the Deep Neural Network (DNN) algorithm resulting in a Word Error Rate (WER) in speech recognition which showed less than optimal results.

Also, Lee and Kwak [7] used DNN and two classifications in detecting sounds based on gender. The two classifications are SVM and decision tree (DT). In his research, feature extraction used by MFCC to identify gender voices resulted in fairly good accuracy.

III. DEEP LEARNING

Deep learning is a method that is often used in the field of machine learning based on the Network Artificial Neural (ANN) principle model. Deep learning can solve problems with large datasets such as image recognition, text detection, speech recognition, audio, etc. Because there are techniques for using feature extraction from training data especially for speech recognition. Artificial Neural Network, a method that adds a hidden layer, this deep learning can be started with the input layer (voice recording), which can then be processed in the form of a signal that is interconnected between nodes with each other in processing data and ultimately through the output to accuracy.

One of the deep learning algorithms is called the Deep Neural Network (DNN). DNN is one of the developments of the Artificial Neural Network. The DNN method is capable of performing voice recognition with good results because it can determine the feature extraction in each layer.

ISSN No:-2456-2165

IV. METHODOLOGY

The basic process for speech recognition systems involves recording speech signals as input, preprocessing, extracting certain features, and finally classifying them. This research has 350 speech recording voice data.

A. Deep Neural Network

Deep Neural Network In three hidden layers is used between the input layer and the output layer. The first hidden layer has 256 units, the second hidden layer has 128 units, and the third hidden layer has 64 units. So that the results of the summary of the model using the hard method have the total number of parameters used in the deep neural network is 48.194 or 48%. Shown in fig. 1.

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 256)	6912
dense_2 (Dense)	(None, 128)	32896
dense_3 (Dense)	(None, 64)	8256
dense_4 (Dense)	(None, 2)	130

Total params: 48,194
Trainable params: 48,194
Non-trainable params: 0

Fig 1:- Input & output parameters of each layer

B. Dataset

This study uses a dataset[8] with male and female genders. The total number of data consisted of 350 speakers, each of whom had 190 men and 160 women who were involved in this voice recording. The corresponding audio is saved as a mono, 16bit, 32kHz WAV file.

C. Extraction Feature

The voice recognition that we have previously processed is then extracted by several methods including spectral centroid, spectral bandwidth, spectral rollof, MFCC (Mel Frequency Cepstral Coefficients), zero-crossing rate, and feature chroma. But here we are focusing on feature extraction for speech recognition using MFCC.

1) Spectral Centroid

Spectral centroid shows the energy frequency spectrum indicating where the center of mass for the sound is located. Shown the fig. 2:

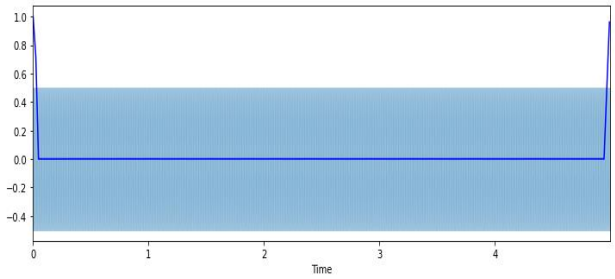

Fig 2:- Centroid Spectral Features

2) Spectral Bandwidth

Spectral Bandwidth of the wave which is the maximum half of the crest represented by two vertical red lines and the wavelength axis. The fig shows. 3:

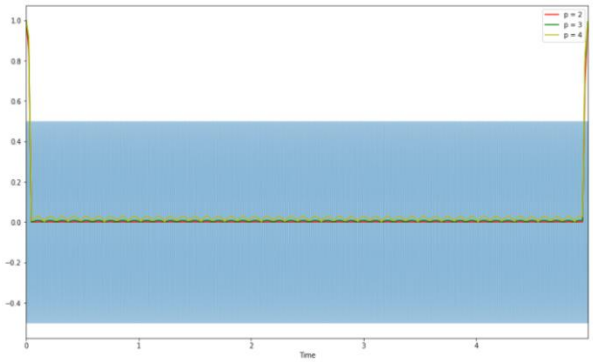

Fig 3:-Bandwidth Spectral Features

3) Spectral Rollof

Spectral rollof to represent the frequency where the high frequency drops to 0. A power spectrum where 85% of the power is at the lower frequency. Fig4 shows:

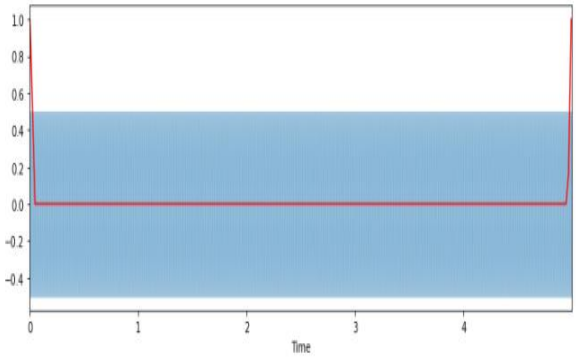

Fig 4:-Rollof Spectral Features

4) Zero-Crossing Rate

Zero-Crossing Rate to measure the smoothness of the signal is to count the number of zero-crossing in the signal segment. Sound signal oscillates slowly ega 100 Hz signal will pass 100 zeros per second. Can be shown the fig. 5:

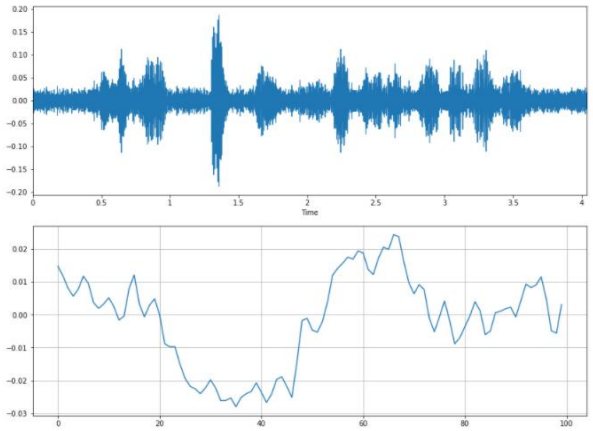

Fig 5:-Zero-Crossing Rate Feature

ISSN No:-2456-2165

5) Chroma Feature

Chroma feature is for calculating chromatogramsfrom a waveform or spectrogram. Chroma feature is a strong spectrum for the representation of music audio, voices, etc. Shown in fig. 6:

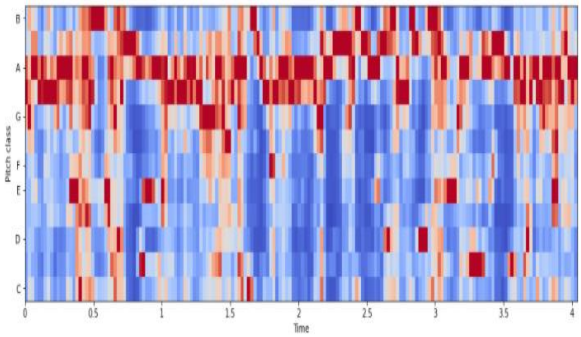

Fig 6:- Chromatogram or Spectrogram Feature

6) MFCC

Mel Frequency Cepstral Coefficients (MFCC) is one method that is widely used in the field of speech technology, both speech recognition, and voice recognition.

The Mel Frequency Cepstral Coefficient (MFCC) technique is often used for the extraction of important features from sound files based on different bandwidth frequency for human hearing, the sound signal is filtered linearly at low frequencies (below 1000Hz) and logarithmically for high frequency (above 1000Hz).

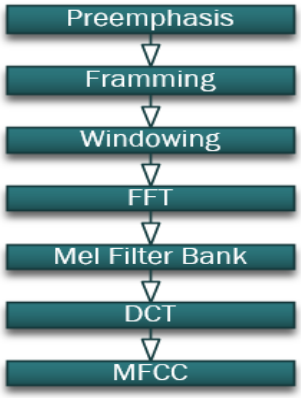

Fig 7:-MFCC Diagram Process

The first diagram process in MFCC is preemphasis, which is to produce energy in high frequency which was previously compressed during the process of producing sound. Framing is used as trimming the sound signal file which is divided into smaller parts, therefore signal analysis can be processed by performing a short time (frame) in the recognition system. So it is most important to cut a signal that is smaller and still contains the original characteristics for the signal analysis process [9]. Windowing is used to avoid interrupting signals that have been previously processed. The function of the hamming window can be expressed in the equation:

$$W[n] = 0.54 - 0.46\ \cos[\frac{2\pi n}{N-1}] \qquad (1)$$

Whereas for the output of each frame after filter processing with N is the number of samples per frame Y [n] is the output signal X [n] and the input signal W [n] is the nth efficiency of the Hamming window [10]. Here is the equation:

$$Y[n] = X[n] * W[n] \qquad (2)$$

Fast Fourier Transform (FFT) is used to convert the signal from time to frequency[11].

The filterbankis based on the Mel scale, namely from linear and logarithmic [12]. The following formula is used to calculate Mel frequency:

$$Mel\,(f) = 2595 *\ \log 10(1 + \frac{f}{700}) \qquad (3)$$

The input from the filterbank of 2595 and 700 is a fixed, predefined value that is widely used in the MFCC method [13]. [14] and the last process is Discrete Cosine Transform (DCT) whose output is called Mel Frequency Cepstral Coefficients (MFCC). Shown in the fig.8:

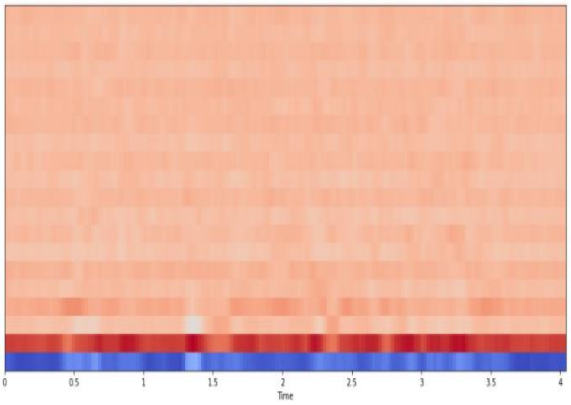

Fig 8:-MFCC Feature

V. EXPERIMENT AND RESULT

The software instrument used to run the system is cloud tools (Google Colab). The reason for using these tools is due to the limited hardware used to support the system processing speed. This section describes the deep neural network for speech recognition classification using SVM and the overall results of speech recognition detection from several matrices in terms of accuracy, precision, recall, and f1. As explained in the following equation:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \qquad (4)$$

$$Precision = \frac{TP}{TP+FP} \qquad (5)$$

$$Recall = \frac{TP}{TP+FN} \qquad (6)$$

$$F1 = \frac{2*(Recall\ x\ Precision)}{(Recall+Precision)} \qquad (7)$$

In the above equation, it is TP true positive, TN true negative, FP false positive, and FN false negative.

Some of these metrics will calculate the success of a prediction whose accuracy results are 0.97%, recall 0.95%, precision 0.88%, and F1 0.96%. Apart from several metrics, a confusion matrix is also presented in this section. The confusion matrix summarizes the classification results to show the performance of the deep neural networks.

A. Support Vector Machine Classification

A support vector machine (SVM) is a conventional learning model for pattern recognition and data analysis. The SVM classification consists of a set of hyperplanes that can be used for classification or regression analysis. After the extraction feature is completewe use the SVM file to measure the accuracy against the voice recognition classification. The results obtained from the SVM classification, the accuracy is 0.9214 or 0.9%. SVM acts as one good approach to data modeling. Also, kernel mapping on vector machines provides a general basis as described in table 1. The highest accuracy that occurs in the RBF kernel is 0.9714% or 0.97%.

Kernel SVM			
	RBF	**Polynomial**	**Linear**
Insample accuracy	0.9785	0.9642	0.9857
Outsample accuracy	0.9714	0.9285	0.9714

Table 1:- RBF,Poly & Linear Kernel Mapping

B. Training and Test Result

In this case, the training was conducted with 80% data sharing for training data and 20% for testing data. As shown in Figs. 9 and 10, is a graph of training accuracy, training loss, validation accuracy, and validation loss. Each value is training accuracy 0.7643, training loss 0.5276, validation accuracy 0.7714, and validation loss 0.5643.

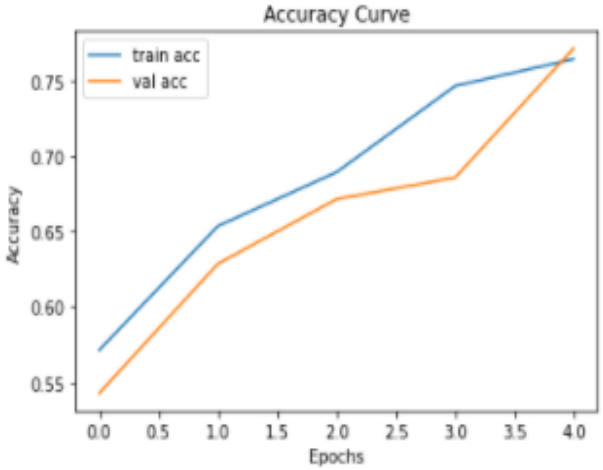

Fig 9:-Accuracy Classification Graph

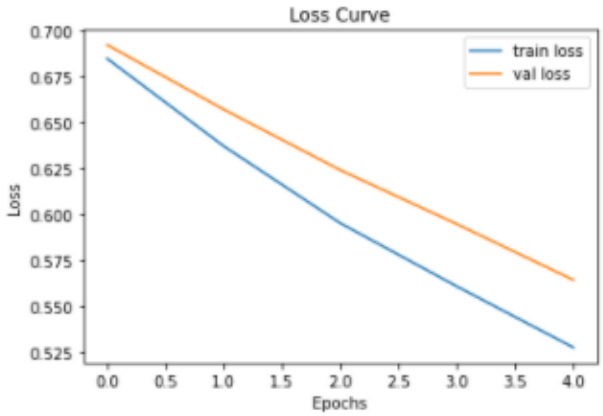

Fig 10:-Loss Classification Graph

This training is carried out after the feature extraction process is tested for its performance stability. This performance stability is measured by the split percentage technique from 60%: 40% to 90%: 10%.

C. Confusion Matrix

This confusion matrix is a representative diagram that is presented with a good process. This confusion matrix summarizes the complete classification results based on true and false objects. The correct classification result is 68 data. Based on Figure 11. The label classification is marked with Female (0) and Male (1).

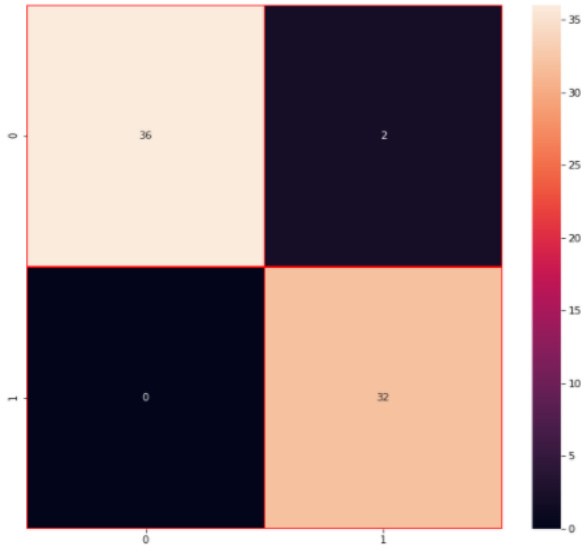

Fig 11:-Confusion Matrix

VI. CONCLUSION

In this study, the detection of voice recognition based on gender used the DNN method which resulted in good accuracy. Also, each speech recognition feature extraction and classification. This extraction feature is to remove noise in gender voice data by using MFCCbecause it is a good, fast, and complete method. Meanwhile, the classification of gender speech recognition uses the SVM algorithm which has good accuracy results.

REFERENCES

[1] Brigham E, The Fast Fourier Transform and Its Application. Prentice-Hall Inc. New Jersey., 1988. .

[2] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. Laurent, Y. Bengio, and A. Courville, “Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks,” In Proceedings of Interspeech, 2016. pp. 410-414.

[3] Jamil Ahmad, Mustansar Fiaz, Soon-il Kwon, Maleerat Sodanil, Bay Vo, and Sung Wook Baik. Gender Identification using MFCC for Telephone Applications - a Comparative Study. arXiv preprint arXiv:1601.01577, 2016..

[4] Martin, A., and Przybocki, M. Speaker recognition in a multi-speaker environment.In Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001) (Aalborg,Denmark, 2001), pp. 787–90. .

[5] Tomi Kinnunen “Spectral Feature for Automatic Voice-independent Speaker Recognition ”Depertment of Computer Science, Joensuu University, Finland. December 21, 2003..

[6] L. Yufang, S. Jie, H. Wenjun, and P. Wenlin, "Speech recognition of isolated word speech of primi language on HTK," Journal of Yunnan Minzu University (Natural Sciences Edition), pp. 426-430, 2015..

[7] M.-W. Lee and K.-C. Kwak, “Performance comparison of gender and age group recognition for human-robot interaction,” IJACSA) International Journal of Advanced Computer Science and Applications, vol. 3, no. 12, 2012 .

[8] Akshay Babbar, "Akshay Babbar : speakerrecognition,"2019. [Online]. Available: https://www.kaggle.com/akshay4/speakerrecognition. [Accessed: 19-Agustus-2020]., [Online].

[9] W. Junqin and Y. Junjun, “An improved arithmetic of MFCC in speech recognition system,” in Electronics, Communications and Control (ICECC), 2011 International Conference on, 2011, pp. 719–722. .

[10] B. J. Mohan, “Speech recognition using MFCC and DTW,” in Advances in Electrical Engineering (ICAEE), 2014 International Conference on, 2014, pp. 1–4. .

[11] A. Vijayan, B. M. Mathai, K. Valsalan, R. R. Johnson, L. R. Mathew, and K. Gopakumar, “Throat microphone speech recognition using mfcc,” in Networks & Advances in Computational Technologies (NetACT), 2017 International Conference on, 2017, pp. 392–395..

[12] N. Alcaraz Meseguer, “Speech analysis for automatic speech recognition,” Institutt for elektronikk og telekommunikasjon, 2009. .

[13] S. T. Saste and S. M. Jagdale, “Emotion recognition from speech using MFCC and DWT for security system,” in Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of, 2017, vol. 1, pp. 701–704..

[14] G. Jhawar, P. Nagraj, and P. Mahalakshmi, “Speech disorder recognition using MFCC,” in Communication and Signal Processing (ICCSP), 2016 International Conference on, 2016, pp. 0246–0250.