IMPROVED LIP-READING LANGUAGE USING GATED RECURRENT UNITS

Nafa Zulfa[1)], Nanik Suciati[2)], and Shintami C. Hidayati[3)]
[1, 2, 3)] Department of Informatics, Faculty of Electrical Technology and Intelligence Informatics,
Institut Teknologi Sepuluh Nopember
Sukolilo, Surabaya
e-mail: naffzzz@gmail.com[1)], nanik@if.its.ac.id[2)], shintami@its.ac.id[3)]

ABSTRACT

Lip-reading is one of the most challenging studies in computer vision. This is because lip-reading requires a large amount of training data, high computation time and power, and word length variation. Currently, the previous methods, such as Mel Frequency Cepstrum Coefficients (MFCC) with Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) with LSTM, still obtain low accuracy or long-time consumption because they use LSTM. In this study, we solve this problem using a novel approach with high accuracy and low time consumption. In particular, we propose to develop lip language reading by utilizing face detection, lip detection, filtering the amount of data to avoid overfitting due to data imbalance, image extraction based on CNN, voice extraction based on MFCC, and training model using LSTM and Gated Recurrent Units (GRU). Experiments on the Lip Reading Sentences dataset show that our proposed framework obtained higher accuracy when the input array dimension is deep and lower time consumption compared to the state-of-the-art.

***Keywords**:Lip-reading, Convolutional Neural Network, Mel Frequency Cepstrum Coefficients, Long Short-Term Memory, Gated Recurrent Units.*

PENGEMBANGAN PEMBACAAN BAHASA BIBIR MENGGUNAKAN GATED RECURRENT UNITS

Nafa Zulfa[1)], Nanik Suciati[2)], dan Shintami C. Hidayati[3)]
[1, 2, 3)] Departementen Informatika, Fakultas Teknologi Elektro dan Informatika Cerdas,
Institut Teknologi Sepuluh Nopember
Sukolilo, Surabaya
e-mail: naffzzz@gmail.com[1)], nanik@if.its.ac.id[2)], shintami@its.ac.id[3)]

ABSTRAK

Pembacaan bahasa bibir merupakan salah satu penelitian dengan tantangan tersendiri dalam visi komputer. Hal tersebut karena pembacaan bahasa bibir membutuhkan banyak data training, waktu dan kekuatan komputasi yang tinggi, dan perbedaan panjang kata. Saat ini, metode penelitian sebelumnya masih mendapatkan akurasi yang rendah dan time consumption yang tinggi, seperti Mel Frequency Cepstrum Coefficients (MFCC) dengan Long Short-Term Memory (LSTM) dan Convolutional Neural Network (CNN) dengan LSTM. Pada penelitian ini, kami menyelesaikannya menggunakan pendekatan yang menghasilkan akurasi tinggi dan time consumption yang rendah. Secara khusus, kami menawarkan pembecaan bahasa bibir menggunakan deteksi wajah, deteksi bibir, filter jumlah data untuk menghindari terjadinya overfitting karena ketidakseimbangan data, ekstraksi citra berdasarkan Convolutional Neural Network (CNN), ekstraksi suara berdasarkan Mel Frequency Cepstrum Coefficients (MFCC) dan training menggunakan Long Short-Term Memory (LSTM) serta Gated Recurrent Units (GRU). Eksperimen pada dataset Lip Reading Sentences menunjukkan bahwa metode yang kami usulkan mampu mendapatkan akurasi yang lebih tinggi ketika dimensi input array dalam dan time consumption yang lebih rendah dibanding metode terdahulu.

Keywords: *Pembacaan bahasa bibir, Convolutional Neural Network, Mel Frequency Cepstrum Coefficients, Long Short-Term Memory, Gated Recurrent Units.*

I. INTRODUCTION

The development of detection technology affects many aspects of human life. Currently, there are various types of detection, such as temperature detection, location detection, and visual detection [1][2]–[4]. One of the various types of detection which have massive research is visual detection. For example, several types of visual detection are facial recognition, object detection, and lip language reading [1], [5], [6]. One of the

factors that influence visual detection is the development of ultraviolet light capture technology. In general, ultraviolet light capture is done using a camera. The camera has various levels in capturing an object symbolized by using the unit of MegaPixel (MP). The larger the MP in a camera, the better the pixel density will be. The pixel density of the camera image affects the accuracy of the detection of visual objects. One of many visual objects detection which has high research demanding is lip reading.

Lip reading is a supporting technology to find out someone's speech based only on visual information given by the shapes formed by the lips and tongue [7]. The lip-reading language aims to help someone communicate and capture the message conveyed through the video, especially people who have hearing impairment and unnative speaker [8]. In addition, lip-reading language will help someone who is not a native speaker capture the message conveyed through the video. many researchers have tried to develop lip reading [7], [9]–[11]. Currently, several study tried to develop lip reading, however the results still are not optimal and require a long training time. Currently, one of the outstanding result obtained propose by Chung [7]. The research [7] proposed lip language reading using a combination of Convolutional Neural Network (CNN), Mel Frequency Cepstrum Coefficients (MFCC), and Long Short-Term Memory (LSTM). The use of CNN and MFCC in feature extraction aims to reduce the loss of important data compared to manual feature extraction. However, the disadvantage of the LSTM is that it has three gates, so it requires a long training process. LSTM also affects system performance because it often causes changes in data information in the stored model so that the resulting accuracy will also below. Therefore, lip-reading language requires a method which can change LSTM to obtain higher accuracy and lower time consumption.

The research conducted by Zaman et al. [12] compare the performance between LSTM and Gated Recurrent Units (GRU) in the prediction of Remo dance movement. The GRU result always gets lower time consumption compared to LSTM. It is because GRU only has two gates.

Therefore, in this study, we propose to use CNN and MFCC to extract the visual and audio information. GRU used to classify the extracted information. Lip reading sentences dataset is utilize in this work [7]. This is because lip reading sentences dataset has inhomogeneous dataset, including light intensity and viewing angles.

II. LITERATURE STUDY

Lip language reading is one of the studies that is currently being intensively developed. However, lip-reading has its challenges, such as camera angle, light intensity, number of frames per second, and video resolution. On the way to answer the state of the art challenges, various methods have been developed, such as lip language reading using (1) CNN, MFCC, and LSTM, (2) CNN and GRU, (3) Facial Landmark Detection and AdaBoost, (4) Facial Landmark Detection and Linear Discriminant Analysis Classifier (LDA), (5) Connectionist Temporal Classification (CTC) and sequence-to-sequence, (6) Color Transformation, Moore Neighborhood Tracing Algorithm and linear SVM classifier, and (7) lip Contour [6], [9]–[11], [13], [14].

CNN and MFCC in previous studies [7] were used to extract the image and audio features. At the same time, LSTM is used to classify lip language. The use of LSTM in the previous method has a drawback, namely, using three gates. The use of many gates will increase the training time. In addition, the use of many gates will also result in frequent changes to the data in the trained model. It is in line with research conducted by Mudaliar [9]. The classification in this research uses LSTM so that it can produce high accuracy. However, in this study, the training time was still relatively long. Meanwhile, in a study conducted by Thabet [10], feature extraction was performed using facial landmark detection. In this study, classification was carried out using AdaBoost and LDA.

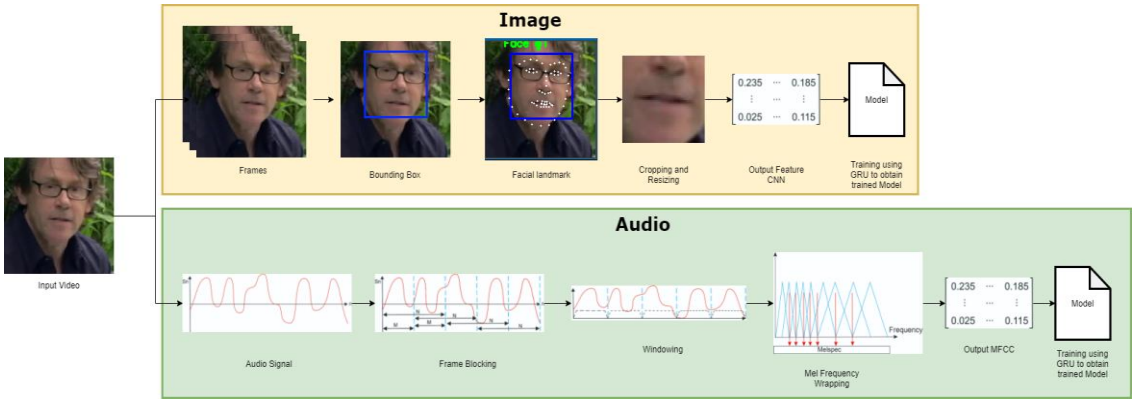

Figure 1. Overall System Framework.

Otherwise, Afouras [6] uses Connectionist Temporal Classification (CTC) and sequence-to-sequence. Meantime, Thein uses [13] Color Transformation, Moore Neighborhood Tracing Algorithm and Linear SVM to better perform. On the other side, Lin [14] uses lip Contour and her algorithm to match to their database to get a better result. Based on several studies above, the use of different methods can produce different accuracy and training times.

III. Overview of The Proposed Method

Figure 1 shows a framework of the lip-reading system. The video input will be pre-processed into image and audio, followed by feature extraction and model training. The result of the training is a ready-to-use model. There are two main steps of the proposed method. It is training process and testing process. The detail of abovementioned step mentioned follows.

A. Training Process

The training process is divided into two process types, namely image and audio. The detail of abovementioned process mentioned follows.

1) Image

The video data input will be extracted into two types, namely image, and sound. Before the video is extracted, the video was segmented for each word is based on the dataset's records. The system detects the bounding box after the video is cut off for each word. In this process, HOG-SVM is used as a bounding box detector [7]. Histograms work by compressing information over a certain sized area. The result of the HOG-SVM process is a bounding box on the face.

Figure 2. Cropped Image Examples.

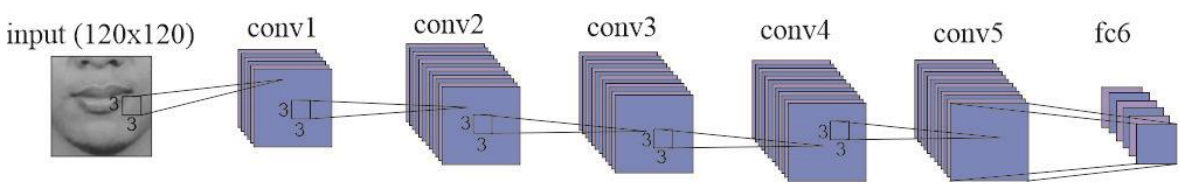

Figure 3. Image Feature Extraction Architecture.

The system will detect facial points in each frame using the facial landmark algorithm after the bounding box on the face is determined. The method on the facial landmark algorithm used is an ensemble of regression trees [15]. One of the libraries that have implemented the HOG-SVM method and an ensemble of regression trees is dlib [16]. After the facial point in each frame has been determined, the image will be cropped in the area around the mouth. Cropping in the area around the mouth is square. The first step in cropping is to take the maximum x-axis and the minimum x-axis at the mouth horizontally. The maximum x-axis result is added to a value of 10, while the minimum x-axis result will be reduced by 10. This is so that the area taken is not too close to the mouth. Meanwhile, the maximum y-axis is obtained by rotating 90 degrees to the maximum x-axis, which has been added a value of 10. Furthermore, the minimum y-axis is obtained by rotating 90 degrees to the minimum x-axis, reduced by 10. The cropped image examples can be seen in Figure 2. The cropping results will then experience feature extraction.

The input data in the form of cropping and resizing results will be processed using CNN. The CNN algorithm will automatically extract the features contained in each frame. Feature extraction on the image is carried out using the VGG-M architecture. The use of the VGG-M architecture is able to produce good and efficient performance [10]. The feature extraction architecture in the figure is shown in Figure 3.

A 3x3 kernel size will be used for the convolutional layer and the pooling layer in the image feature extraction. At the same time, the input on the input layer is a frame from the lips that were previously cropped and resized. To get the features of each frame, the process will continue on the convolutional layer. The result of the convolutional layer will experience an activation function called ReLu. After experiencing the activation function, the activation results will enter the pooling layer. The iteration process using the convolutional layer, activation function, and pooling layer will be repeated five times to get maximum results. The final step of the CNN is that the results of the iteration five times will enter the fully connected layer. The result of the extraction process using CNN is a unique feature found in each frame.

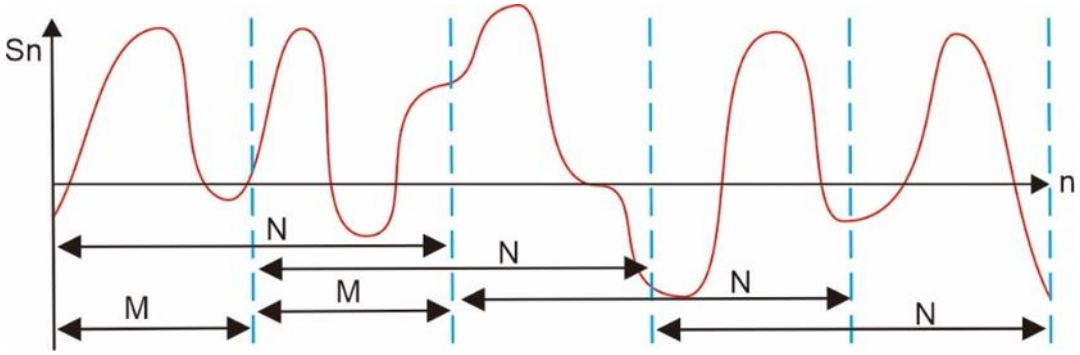

Figure 4. Audio Feature Extraction Architecture.

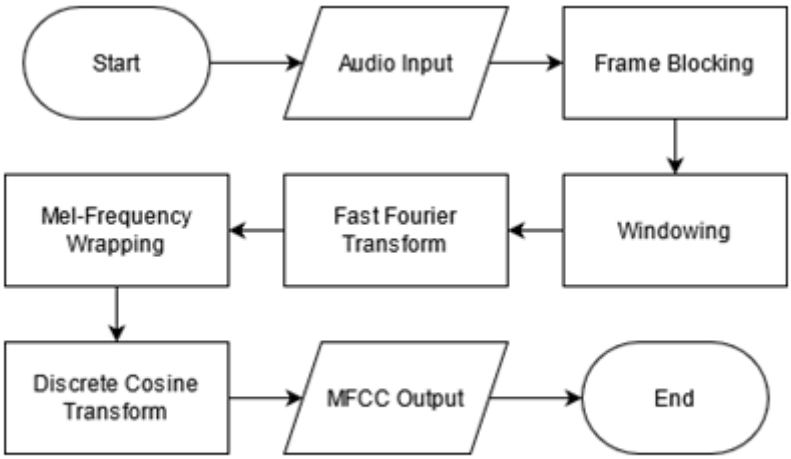

Figure 5. Audio Feature Extraction Flowchart.

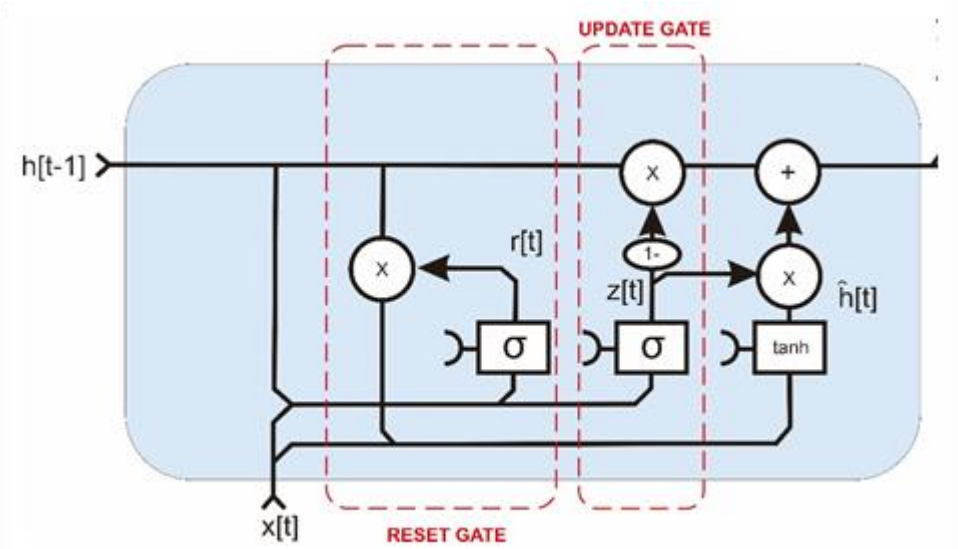

Figure 6. GRU Architecture.

Figure 7. Dataset Lip Reading Sentences Video Samples.

The result of the sound feature and image feature will be entered into the GRU. The GRU architecture can be seen in Figure 7. The first step in the GRU is to enter new information through the reset gate. The reset gate is used to determine whether the newly stored information data will be discarded. The results of the determination will be done pointwise multiplication with the previous hidden state layer. In addition, new information is also entered into the update gate. Update gate serves to determine whether the stored information data will be updated.

The result of the update gate will be done pointwise multiplication with the previous hidden state layer. The results of the pointwise multiplication will be carried out pointwise addition with the results on the reset gate that has undergone activation function and pointwise multidimensional with the update gate. The final result of the GRU is a new hidden state. Many new hidden states will be saved as model.

2) Audio

The segmented video converted into audio. The conversion of audio format uses WAV (Waveform Audio Format) because it has less compression quality compared to the format of the other. The audio results will immediately experience feature extraction using MFCC. The audio feature extraction flowchart can be seen in Figure 5.

The first step in feature extraction using MFCC is to input sound into the system. The input sound will be frame blocking. At this stage, the voice signal consisting of S samples is divided into several frames containing N samples. Finally, each frame is separated by M. The process at this stage can be seen in Figure 4. Windowing is performed for each frame of the voice signal. It aims to minimize signal discontinuities at the beginning and end of each frame. The concept of applying windowing by tapering the signal to zero at the beginning and end of each frame.

Fast Fourier Transform FFT functions to convert each frame containing N samples from the time domain to the frequency domain. The next stage is Mel-Frequency Wrapping. In Mel-Frequency Wrapping, the FFT signal is grouped into a triangular filter file. The triangular filters at this stage overlap each other. At this stage too, each FFT value is multiplied by the corresponding filter gain. Then the multiplication results will be added up. The goal is Mel-Frequency Wrapping so that each group contains a signal energy weight. The result of the Mel-Frequency Wrapping stage is that each spectrum has a signal energy weight.

The last step after getting the signal energy weight of each spectrum is to convert the spectrum weight into a cepstrum using Discrete Cosine Transform (DCT). The entire process in the training section will be repeated for testing, except train using GRU process. The sound feature and image feature result will be entered into GRU. GRU will be classified the calculation result based on the trained model.

IV. RESULT AND DISCUSSION

In this section we describe dataset, experimental configuration, and comparison with different combinations of the proposed method. The detail of abovementioned section mentioned follows.

A. Dataset

The data used in this study is a lip-reading sentences dataset from the BBC [7]. This dataset consists of 41,427-word classes with data in the form of video. The word class in the dataset is the type of word spoken in the video. The dataset contains news videos, both from interviews with sources and journalists reading the news. The videos on the dataset are in English. Video on the dataset has frames per second (fps) of 30 fps. The dataset contains more than 10,000 video data. Each video is less than 10 seconds long. The examples of a video can be seen in Figure 7.

B. Experimental Configuration

This research experiment uses Intel Core i7 10700K with 128GB DDR4 3200Mhz Memory and RTX 3080 10GB. This experiment uses python programming language with an anaconda environment. The experiment scenario uses sound and image cases. The research experiment scenarios are shown in detail in Table 1. All experiment scenarios apply k-fold cross-validation with k=10. It aims to avoid overfitting and underfitting. Each test scenario uses a batch size of 256 with 25 epochs and uses the Adam optimizer with a learning rate of 0.0001.

The experiment was carried out with two filter conditions, namely when each word class had a minimum of 10,000 and 15,000 data. The filtering process will affect the number of word classes that meet the requirements so that several conditions are created, such as 11, 17, 26, and 68-word classes. At the same time, the type of data in this experiment is divided into two types, namely images and sound. The type of data available affects the feature extraction method that can be applied. Feature extraction on image data uses CNN, while voice data uses MFCC and its derivatives.

The MFCC and CNN result will be padded with 0 to create the exact length duration for each feature map. The sound feature map padded into 317xN dimension. N is the deeper MFCC value and the value of its derivative. On the other hand, the images feature maps padded into 185x512 dimensions. The result of the padded feature map was classified using LSTM and GRU.

The experiments are based on accuracy, error rate and time consumption. The accuracy is calculated using Equation (1).

$$Accuracy = \frac{\sum C}{\sum A} * 100\%, \quad (1)$$

$\sum C$ is the number of cases correctly identified. $\sum A$ is the number of all prediction. On the other hand, the error rate is calculated using Character Error Rate (CER). The formula of CER is showed in Equation (2).

$$CER = \frac{(Ic+Sc+Dc)}{(Ic+Sc+Dc)}, \quad (2)$$

C. Sc *is the changed character count.* Dc *is the deleted character count.* Ic *is the inserted character count. However,* Cc *is the correct character count. Morevoer, time consumption is calculated from the system load the image until the training process is done. Comparison with Different Combinations of The Proposed Method*

Table 2 shows the experimental results using the Lip Reading Sentences dataset cut off for each word. The experiment uses a filter size of 5000 and 10000. The filter shows the minimum amount of data in each word class. The experimental results show that the more data used, the more accuracy and lower the Character Error Rate (CER); however, the deeper the array dimensions used are not directly proportional to the accuracy and the resulting CER. The deeper array dimensions used affects the accuracy of GRU, which is higher than LSTM. LSTM accuracy could not be as good as GRU when on deeper data because the calculations on the LSTM are getting more complex, so the addition of a one-dimensional feature map will significantly affect the results. The highest accuracy and the lowest CER are shown when using the audio case with a combination of MFCC Level 4 and GRU of 74.95% and 36.95%. This is because the GRU has fewer gates than the LSTM, so the changes in the data values in the model are less. The lowest accuracy is shown when using the image case. It happens because the heterogeneity, the low video resolution, and the noise of video data are too high so that the pattern determined by the architecture offered is still low. Furthermore, audio and video classification result could not be optimal because the amount of data in each class is imbalance. Moreover, the duration between each video word is different. However, the result of audio and video classification could not get optimal; the time consumption generated when using GRU is always lower than LSTM. This is because the GRU architecture only has two gates, thus allowing the use of less computational time than LSTM, which has three gates.

TABLE 1
EXPERIMENT SCENARIO DETAIL.

Data Type	Combination	Filter	Class Total
Image	CNN - LSTM	10,000	21
Image	CNN - LSTM	15,000	21
Image	CNN – GRU	10,000	11
Image	CNN – GRU	15,000	11
Audio	MFCC – LSTM	10,000	68
Audio	MFCC – LSTM	15,000	17
Audio	MFCC Level 2 – LSTM	10,000	26
Audio	MFCC Level 2 – LSTM	15,000	17
Audio	MFCC Level 4 – LSTM	10,000	26
Audio	MFCC Level 4 – LSTM	15,000	17
Audio	MFCC Level 6 – LSTM	10,000	26
Audio	MFCC Level 6 – LSTM	15,000	17
Audio	MFCC – GRU	10,000	26
Audio	MFCC - GRU	15,000	17
Audio	MFCC Level 2 – GRU	10,000	26
Audio	MFCC Level 2 - GRU	15,000	17
Audio	MFCC Level 4 – GRU	10,000	26
Audio	MFCC Level 4 - GRU	15,000	17
Audio	MFCC Level 6 – GRU	10,000	26

TABLE 2
EXPERIMENT SCENARIO DETAIL.

Combination	Filter	Padding	Accuracy	Time Consumption (Seconds)	CER
CNN - LSTM	10,000	185	4.76%	589648.08	93.70%
CNN - LSTM	15,000	185	9.09%	582413.80	96.50%
CNN – GRU	10,000	185	4.76%	588297.50	96.67%
CNN – GRU	15,000	185	9.09%	580531.19	98.40%
MFCC – LSTM	10,000	317	61.60%	162440.86	37.61%
MFCC – LSTM	15,000	317	69.72%	163118.94	31.16%
MFCC Level 2 – LSTM	10,000	317	67.66%	140522.52	31.07%
MFCC Level 2 – LSTM	15,000	317	74.51%	140716.93	25.68%
MFCC Level 4 – LSTM	10,000	317	68.65%	149247.70	41.64%
MFCC Level 4 – LSTM	15,000	317	74.87%	148902.65	37.17%
MFCC Level 6 – LSTM	10,000	317	67.32%	158480.69	50.95%
MFCC Level 6 – LSTM	15,000	317	73.75%	157880.78	48.73%
MFCC – GRU	10,000	317	61.58%	161915.46	38.09%
MFCC - GRU	15,000	317	68.57%	163005.04	31.91%
MFCC Level 2 – GRU	10,000	317	67.15%	140140.59	31.21%
MFCC Level 2 - GRU	15,000	317	74.42%	140192.35	25.86%
MFCC Level 4 – GRU	10,000	317	69.72%	148963.52	41.43%
MFCC Level 4 - GRU	15,000	317	74.95%	148504.08	36.95%
MFCC Level 6 – GRU	10,000	317	68.05%	157958.30	50.71%

V. CONCLUSION

This paper offers an increase in lip-reading with the Lip Reading Sentences dataset using the GRU classifier. The input data extracts into two categories: image and audio. Image data extracted using CNN based on VGG-M architecture and classified using LSTM and GRU to get the best possible result. At the same time, audio data extracted using MFCC, MFCC Level 2, MFCC Level 4, and MFCC Level 6. The result of MFCC and its heredity will be classified using LSTM and GRU to get the optimize result. The results obtained based on accuracy, time consumption, and error rate. The purpose of this study is to increase accuracy, reduce time consumption and the resulting error rate. The proposed system can achieve the main goal by replacing the existing classifier with GRU. As for the voice scenario, an increase occurs when using MFCC Level 4. The research offered can be used to be applied in real-time so that it can help a deaf and non-native speaker to understand the speech of the other person.

REFERENCES

[1] A. L. Akbar, C. Fatichah, and A. Saikhu, "Pengenalan Wajah Menggunakan Metode Deep Neural Networks Dengan Perpaduan Metode Discrete Wavelet Transform, Stationary Wavelet Transform, Dan Discrete Cosine Transform," *JUTI J. Ilm. Teknol. Inf.*, vol. 18, no. 2, pp. 158–170, 2020.

[2] C. Pandian *et al.*, "Raman distributed sensor system for temperature monitoring and leak detection in sodium circuits of FBR," *ANIMMA 2009 - 2009 1st Int. Conf. Adv. Nucl. Instrumentation, Meas. Methods their Appl.*, pp. 8–11, 2009, doi: 10.1109/ANIMMA.2009.5503761.

[3] M. T. Kian and L. L. Choi, "GPS and UWB integration for indoor positioning," *2007 6th Int. Conf. Information, Commun. Signal Process. ICICS*, 2007, doi: 10.1109/ICICS.2007.4449630.

[4] H. J. Seo and P. Milanfar, "Visual saliency for automatic target detection, boundary detection, and image quality assessment," in *Image Processing*, 2010, pp. 5578–5581, doi: 10.1109/ICASSP.2010.5495239.

[5] A. Saini and M. Biswas, "Object detection in underwater image by detecting edges using adaptive thresholding," *Proc. Int. Conf. Trends Electron. Informatics, ICOEI 2019*, no. Icoei, pp. 628–632, 2019, doi: 10.1109/ICOEI.2019.8862794.

[6] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep Audio-visual Speech Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–13, 2018, doi: 10.1109/TPAMI.2018.2889052.

[7] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 3444–3450, 2017, doi: 10.1109/CVPR.2017.367.

[8] E. B. Nitchie, *Lip-Reading Principles and Practise*, vol. 11. Frederick A. Stokes Company, 2008.

[9] N. K. Mudaliar, K. Hegde, A. Ramesh, and V. Patil, "Visual Speech Recognition: A Deep Learning Approach," no. Icces, pp. 1218–1221, 2020, doi: 10.1109/icces48766.2020.9137926.

[10] Z. Thabet, A. Nabih, K. Azmi, Y. Samy, G. Khoriba, and M. Elshehaly, "Lipreading using a comparative machine learning approach," *Proc. IWDRL 2018 2018 1st Int. Work. Deep Represent. Learn.*, pp. 19–25, 2018, doi: 10.1109/IWDRL.2018.8358210.

[11] J. S. Chung and A. Zisserman, "Lip reading in the wild," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10112 LNCS, pp. 87–103, 2017, doi: 10.1007/978-3-319-54184-6_6.

[12] L. Zaman, S. Sumpeno, and M. Hariadi, "Analisis Kinerja LSTM dan GRU sebagai Model Generatif untuk Tari Remo," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 8, no. 2, p. 142, 2019, doi: 10.22146/jnteti.v8i2.503.

[13] T. Thein and K. M. San, "Lip movements recognition towards an automatic lip reading system for Myanmar consonants," *Proc. - Int. Conf. Res. Challenges Inf. Sci.*, vol. 2018-May, no. 1, pp. 1–6, 2018, doi: 10.1109/RCIS.2018.8406660.

[14] B. S. Lin, Y. H. Yao, C. F. Liu, C. F. Lien, and B. S. Lin, “Development of novel lip-reading recognition algorithm,” *IEEE Access*, vol. 5, no. c, pp. 794–801, 2017, doi: 10.1109/ACCESS.2017.2649838.
[15] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, no. August, pp. 1867–1874, 2014, doi: 10.1109/CVPR.2014.241.
[16] D. King, "pypi," 5 12 2020. [Online]. Available: https://pypi.org/project/dlib/. [Accessed 19 1 2021].