Facial Movement Recognition Using CNN-BiLSTM in Vowel for Bahasa Indonesia

Muhammad Daffa Abiyyu Rahman
Departement of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
6022211015@mhs.its.ac.id

Alif Aditya Wicaksono
Departement of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
aditya.18072@student.its.ac.id

Eko Mulyanto Yuniarno
Departement of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ekomulyanto@ee.its.ac.id

Supeno Mardi Susiki Nugroho
Departement of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
mardi@its.ac.id

***Abstract*—Speaking is a multimodal phenomenon that has both verbal and non-verbal cues. One of the non-verbal cues in speaking is the facial movement of the subject, which can be used to find the letter being spoken by the subject. Previous research has been done to prove that lip movement can translate to vowels for Bahasa Indonesia, but detecting the whole facial movement is yet to be covered. This research aimed to establish a CNN-BiLSTM model that can learn spoken vowels by reading the subject's facial movements. The CNN-BiLSTM model yielded a 98.66% validation accuracy, with over 94% accuracy for all five vowels. The model is also capable of recognizing whether the subject is currently silent or speaking a vowel with 98.07% accuracy.**

***Keywords*— bahasa indonesia vowel, BILSTM, CNN, face movement, recognition**

I. INTRODUCTION

Speaking is a multimodal phenomenon that has both verbal and non-verbal cues [1]. As a verbal-based communication, the speaking activity generated sound signals that can be analyzed. As such, much research that covers understanding a dialogue or monologue utilizes the signal that is generated by this activity [2]. However, this approach can be challenging when the sound data is inadequate due to issues such as ambient noise, low Signal-to-Noise Ratio (SNR), and microphone distance [3]. On the other hand, the non-verbal cues, such as the visual cue of the speaker, are yet to be deeply explored.

The growth of computer vision field allowed better analysis of image-related problems, including the problem of non-verbal cues for speaking activity. Prom-on and Onsri discovered that there is a correlation between facial movement and sound acoustic being produced [4], which validates the nature of speaking as a multimodal phenomenon. Another non-verbal option used for determining spoken words is the lip-reading method [5]. These information translated to the potential of utilizing computer vision to assist research related to speaking detection and recognition problems.

One way of utilizing the computer vision field in speaking activity is making visual recognition a complementary aspect of the existing sound signal approach. Kumatani and Stiefelhagen showed in their research with a result that shows the capability of recognizing the speaker's words improves significantly when both sound and visual cues are combined [6]. Another research done by Isobe et al for Japanese language also combined both sound and visual cues [7].

However, this approach still had a dependency on sound data, making it susceptible to sound data issues. An approach that focuses on visual cues become a consideration to minimize reliance on sound data, which translated to reduced susceptibility against sound data issues. This visual focused approach is known as visual-only approach.

Visual-only approach, where spoken word or letter is decided solely based on visual cues, spanned across multiple languages, such as English [8], Japanese [9] and Bahasa Indonesia [10]. For Bahasa Indonesia, by focusing on the nature of lip motion acting as a visual cue of the speaker, Maxalmina concluded that it is possible to detect specific vowels with the highest accuracy of 84%.

Maxalmina's lip motion approach shows that visual cues can be as useful to determine specific vowels that are spoken by a subject. However, the selection of lip motion as the approach rendered information that might be present outside of the lip area, such as the cheek, jaw, and other areas of the face, not covered.

This research attempted to create a model to detect spoken vowels by utilizing face movement, including but not limited to the cheek and jaw areas of the face. The model created with this research is aimed to attain an accuracy better than Maxalmina's lip motion.

II. METHODS

An overview of the research method can be found in Fig 1. The Input for this research is dataset of images. The images

How to cite: Rahman, M. D. A., Wicaksono, A. A., Yuniarno, E. M., Nugroho, S. M. S. (2024). Facial Movement Recognition Using CNN-BiLSTM in Vowel for Bahasa Indonesia. *JAREE (Journal on Advanced Research in Electrical Engineering), 8(1).*

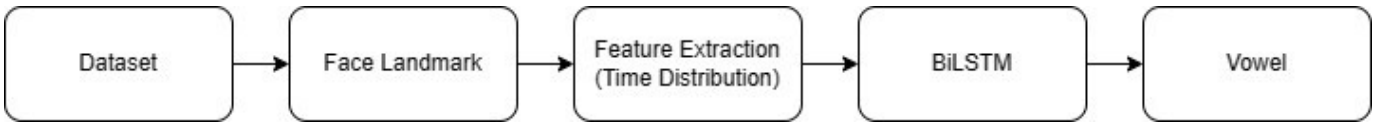

Fig. 1. Research Method

within the dataset is then processed utilizing MediaPipe library [11] to take the face landmarks extracted from the images.

The extracted face landmark features are then processed with a time distribution model to get the key features needed for the next step. BiLSTM is utilized to understand the sequence of key features. These sequences are then categorized into one of the five vowels in Bahasa Indonesia, or a silent state to denote the moment a user is not speaking. Thus, the final output of this method is either one of the vowels or a silent state.

A. Dataset

The dataset for this research is a self-made dataset of images with an image size of $480p$ per image in the form of photo frames. A total of 33,091 images are taken for the dataset, split into 6 categories representing all five vowels and an additional state where the speaker is silent.

The distribution of the image consisted of 4,360 images for letter A, 5,991 images for letter E, 5,169 images for letter I, 5,024 images for letter O, 5,980 images for letter U, and 6,567 images for the silent state. This distribution is represented in table I. The dataset has no augmented data, thus each data is taken in its original form.

To gather all these images, a laptop camera is used to get a video record, configured at $30fps$. The video is then split apart to generate images with 30 images representing a one-second recording of the video. There are a total of 60 videos, 10 for each vowel and 10 for the silent state, each with varying length. These original images are taken at several locations with one subject. All of them are colored with enough contrast to differentiate the subject and the background. In addition, the subject always faces the camera, with the head turned around 45 degrees to the left or right on some images to give the model understanding of the vowel spoken by the subject even when not directly facing the camera. An example of the image is provided in Fig 2.

Speaking is an activity that happens over some time, hence it can be viewed as a time-based problem. Since facial movement is an image problem, Speaking activity can be viewed as both a time and image problem.

To fit the need for a time-based form for the gathered data of images in table I, the data is merged into sets of images for each input where each of the image sets consisted of 30 images representing 1 second of time based on 30 fps used in the original data recording. The data is shifted every 3 images to show a difference in time, creating small gaps that allow the model to predict what happened between those frames. Table II provides the number of data for each vowel.

TABLE I
FACE DATA FOR EACH VOWEL

Letter	Total Images
A	4,360
E	5,991
I	5,169
O	5,024
U	5,980
-	6,567
Total	33,091

Fig. 2. Example of Image in the Dataset

Since the original data have differences in the number of images available, the CNN-LSTM data is also imbalanced for each letter. The data shrunk by approximately 66% of the original amount due to the frame gap implementation, but it is necessary to note that each data contains 30 images, some of which are the same in a different order in each image set to represent the concept of moving in time.

B. Face Landmark

MediaPipe is utilized to assist in generating the face landmarks from the original image, which is needed for extracting key features to detect the vowel being spoken, or if the user is currently silent or speaking.

To generate the landmark, each image in the set of images on table II is then processed with MediaPipe to generate the landmarks on dark background as shown in Fig 3. The generation process consisted of the subject's face in each image being filtered out from the background, creating information for facial features that represented the subject's face in the original image.

The facial features are then mapped into a pitch-black image by connecting the facial features with white lines. This created a distinction in the resulting image as shown in Fig 3, where

TABLE II
DATA FOR EACH VOWEL IN TIME-BASED FORM

Letter	Total Data
A	1,444
E	1,988
I	1,714
O	1,665
U	1,984
-	2,180
Total	10,975

Fig. 3. Face Landmark Process

the background becomes pitch black while the subject's facial features are colored white, making it easier to distinguish and be processed by the neural network. The image size of the new image is the same as the base image, which is 480*p*. A time-distributed result of this process is utilized for the next step, which is the feature extraction and selection.

C. Feature Extraction and Selection (Time Distribution)

The feature extraction and selection are done with a time-distributed convolution model provided in Fig 4. The landmark images are used as the input for feature extraction and selection.

There are four steps to select the best features, consisting of a 2D convolution, max-pooling, flattening, and then selection. The 2D convolution process extracted features from the image. Max-pooling then finds the best of these extracted, which are then flattened for selecting specific features for future predictions.

The time distributed aspect is related to speaking as an event that happened over some time, hence there will be multiple face landmarks that received feature extraction and selection for the next step.

D. BiLSTM

Speaking is an activity that happened over some time, thus a form of neural network with back-propagation capability [12] is needed to allow the network to understand the information within a set of time. This opened a lot of options, such as Recurrent Neural Network (RNN) that has vanishing gradient problem [13], or the Long Short-Term Memory [14] which fixed the issue, but limited to remembering one-way sequences.

Bidirectional LSTM, shortened as BiLSTM, has the advantage of understanding two-way sequences, while also averting the dangerous vanishing gradient problem. This helped the model to understand the whole sequence better for determining the vowel or silent state of the current sequence better than previous models.

Fig 5 provided the BiLSTM process for this research. The result of feature extraction becomes the input of the BiLSTM, where the BiLSTM then processed the input into a sequence as the output that will be utilized for prediction as shown in Fig 6.

The vowel prediction is selected with a Sigmoid activation. Selection of Sigmoid opposed to Softmax is due to an issue with bottlenecks within Softmax [15].

E. Vowels

Vowels are voiced, central-oral friction-less sound when defined in a purely phonetic way. For linguists, the phonetic definition is considered complicated and never an exact correspondence with how some languages have cases of not exactly fulfilling the criteria that go with the basic definition provided [16]. For the specific case of Bahasa Indonesia, the definition is considered sufficient to define the vowels.

Bahasa Indonesia vowels consisted of five letters. They are A, E, I, O, and U. Vowel E has two diacritics according to the recent Ejaan Yang Disempurnakan V, shortened as EYD V, convention [17]. Table III provided an example of each vowel at the start, middle, and end of a word according to the new convention, with vowel E having two lines, each line for one diacritic.

EYD V convention overrides the previous convention of Pedoman Umum Ejaan Bahasa Indonesia, shortened as PUEBI, as of 16 August 2022 [18]. The difference between both conventions for the vowel system lies in the interpretation of the letter E. The previous PUEBI convention stated vowel E had 3 diacritics, while the current EYD V stated vowel E only had 2 diacritics. However, this research, as well as Maxalmina's research [10], does not account for the distinction of vowel E as a focus and treated all diacritics of vowel E as one vowel.

Since vowels come as part of the words spoken by the subject, the subject might not be speaking and remain silent. This state, if not defined, is capable of causing a misinterpretation of the vowel currently spoken by the subject. To avoid such misinterpretation, the silent state must be considered as part of the final output.

As such, the vowels and a silent state are used as the final output from the whole process. This process will determine based on the previous processes whether the original input, which is a set of images, corresponded to one of the five vowels or if the subject is silent. In the overall design, this translated to the final output having six possibilities, five vowels, and a silent state as shown in Fig 6. For this research, the silent state is denoted as - (en dash).

Fig. 4. Features Extraction and Selection

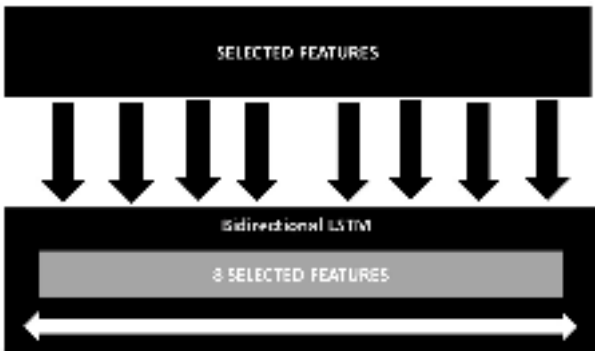

Fig. 5. BiLSTM Input and Process

Fig. 6. Vowel Prediction

TABLE III
BAHASA INDONESIA VOWELS ACCORDING TO EYD V

Vowel	Start of Word	Middle of Word	End of Word
a	api	padi	lusa
e*	enak	petak	sore
	emas	kena	tipe
i	itu	simpan	murni
o	oleh	kota	radio
u	ulang	bumi	ibu

III. RESULTS AND DISCUSSION

A. Model Training and Tuning

The final model for this research is provided by table IV. The model is trained with a distribution of 75% data for training and 25% data for testing using data from table II. There are several design choices for the model, such as LeakyRELU for feature extraction and selection, utilization of the Adam algorithm for the optimizer, Categorical Cross Entropy as the loss function, and the inclusion of a Dropout layer.

The reason for selecting LeakyRELU over RELU for the convolution part lies in the dead neuron problem, part of vanishing gradient problems, where the neuron value becomes zero and rendered inactive [19]. The LeakyRELU solved this issue by adding a small value on inactive state [20].

On the other hand, the Adam algorithm is selected as the optimizer due to the computational efficiency, minimal memory requirements, and considered well-suited for problems that are large in terms of data [21].

Categorical Cross Entropy is selected as the loss function due to the nature of this prediction task being categorical. There are six possible outputs for the final layer, which translated to six possible results. The model assigns a prediction value to each of the six possible results, with the highest value selected as the predicted result.

The dropout layer is included in the model to reduce the possibility of entering an overfitting state [22].

There is a major challenge with the CNN-BiLSTM model provided by table IV. The challenge is that the model is capable of trapping itself in local minima, causing the model's accuracy to stagnate even after several epochs. However, this challenge can be overcome by tuning the model's hyperparameter, specifically the batch size utilized for training.

Table V provided an overview of how the batch size influences the accuracy of both training and validation of the model. When batch size is raised, the level of accuracy increased significantly. The model is trained with a batch of

TABLE IV
FULL MODEL

Layer	Layer Detail		
	Filter/Unit	Kernel	Activation
TimeDistribution Convo 2D	4	5	LeakyRELU
TimeDistribution MaxPool 2D	-	-	-
TimeDistribution Flatten	-	-	-
TimeDistribution Dense	4	-	LeakyRELU
Bidirectional Sequential LSTM	8	-	TanH
Flatten	-	-	-
Dropout (0.2)	-	-	-
Dense	6	-	Sigmoid

2, 8, and 16 batches, with the model's accuracy significantly improved when a higher batch number is used.

When the batch size of 2 is used, the model is trapped within the 14% and 20% accuracy range, preventing the model from improving its prediction capability. This trapped in local minima is sometimes bypassed but created a lot of unreliability in the model training.

A batch size of 8 allowed the model to perform better in escaping the local minima. However, it took quite several epochs to improve the model's accuracy, which improved gradually over time. Within the first 3 epochs, the model only reached 41.15% accuracy and 46.38% validation accuracy. This result, however, showed that raising the number of batches is the right direction to improve the model's performance.

The best result comes from utilizing an epoch of 16, where the model attained an accuracy of 96.26% and validation accuracy of 98.66%, more than double the accuracy provided with a batch size of 8. The result indicated that this batch size is considered most suitable for the proposed model and 16 is selected as the model's batch size.

B. Best Result

The best result, as shown in table V, is the one with a batch size of 16. A more detailed version of this training can be found in table VI which showed that the model gained significant accuracy improvement with each epoch, where the model reached over 90% accuracy by the third epoch and over 90% validation accuracy by the second epoch.

In addition, no mark of overfitting nor underfitting was shown with the validation loss descending faster than the training loss. Accuracy improvement is visible through the growth of both training and validation accuracy, yielding 96.26% and 98.66% respectively for the model at the final epoch.

The model is then tested against the entirety of data from table II. A summary of the result can be found in table VII. From the testing against the entire data, the model is found to be capable of predicting each vowel with an average accuracy of 97.66%.

The model had a hard challenge with vowels A and E, yielding 94.53% and 94.62% accuracy respectively. These two vowels had the highest tendency to be misinterpreted as vowel U, with 4.22% and 4.83% errors for vowels A and E respectively. This showed that the model sometimes confused vowels A, E, and U to a degree, and had a harder time determining the exact vowel spoken by the subject if the subject's facial features is too subtle for the model to differentiate between the three vowels.

On the other hand, the model is capable of differentiating vowel I, O, and U with an accuracy of over 99%, showing an extreme level of accuracy in the detection of these three vowels. The best accuracy, however, is shown in determining vowel I and O, with the model misinterpretation of other vowels as these two, are very low, which translated to the model being truly capable of recognizing these two vowels with the highest accuracy.

In addition, the model is capable of detecting whether the speaker is silent or speaking with an accuracy of 98.07% for data from a silent state. Furthermore, the model also never interpreted a speaking subject as silent, shown by all five vowels having 0% of them interpreted as a silent state by the model.

The results showed that the model knows if the subject is speaking, but can be doubtful if the subject is silent if their facial movement indicated some form of attempt to speak a vowel, shown by the 2.93% total error across all five vowels for the silent state dataset.

Fig 7 showed an example of a sequence when the subject opening their mouth influences the model's silent state interpretation. Within the entirety of the sequence, the subject mouth is opened over several frames, as if the subject is speaking, which caused the model to misinterpret the subject as speaking despite being silent.

C. Comparison with Maxalmina Lip Motion Model

One of the recent research closest to visual cues for Bahasa Indonesia is Maxalmina's 3D CNN Lip Motion. It utilizes the lip motion for vowel prediction, while this research utilized the entire facial movement to predict the vowel. The CNN-BiLSTM model utilized in this research is a contrast to Maxalmina's approach that utilized 3D convolution and relies completely on Convolutional Neural Network as shown in table VIII. A double fully connected layer is utilized to assist in predicting the vowel spoken by the subject after a 3-step 3D convolution is done on the input image for Maxalmina's Lip Motion.

The level of accuracy in Maxalmina's model [10] overall is 84% accuracy across all five vowels. However, that model does favor vowel U with a 9% error, while it has the highest error rate against vowel I with a 20% error. Table IX provides an overview of the error rate comparison of this model and Maxalmina's lip motion model.

The accuracy difference shown in table IX showed that the facial movement recognition approach this model took is advantageous when compared to a lip motion approach. By analyzing additional information provided by the entirety of the speaking subject's face, the accuracy of prediction improved significantly.

This significance can be found with the highest accuracy improved from 84% to 99.94%. In addition, this research

TABLE V
PARAMETER TUNING: BATCH SIZE

batch size	2		8		16	
epoch	accuracy	val accuracy	accuracy	val accuracy	accuracy	val accuracy
0	17.83%	14.82%	19.63%	19.43%	38.16%	53.57%
1	18.24%	19.08%	29.90%	37.34%	73.75%	91.49%
2	17.33%	14.79%	41.15%	46.38%	96.26%	98.66%

TABLE VI
BEST MODEL PERFORMANCE

epoch	accuracy	loss	val accuracy	val loss
0	0.3816	1.5086	0.5356	1.1129
1	0.7375	0.7052	0.9149	0.2732
2	0.9626	0.1321	0.9865	0.0627

TABLE VII
CNN-BILSTM CONFUSION MATRIX

Target	Predicted					
	A	E	I	O	U	-
A	**94.53%**	0.90%	0.35%	0%	4.22%	0%
E	0.50%	**94.62%**	0.05%	0%	4.83%	0%
I	0.06%	0.57%	**99.53%**	0%	0.41%	0%
O	0	0.06%	0	**99.94%**	0%	0%
U	0.51%	0%	0.20%	0%	**99.29%**	0%
-	0.05%	0.46%	0.18%	1.24%	0%	**98.07%**

Fig. 7. Sequence of Silent Subject Misinterpreted as Speaking

TABLE VIII
MAXALMINA'S LIP MOTION MODEL

Layer	Layer Detail		
	Kernel	Output	Activation
Convo 3D (32)	3x3x3	8x54x110x32	Relu
MaxPool 2D	1x2x2	8x27x55x32	-
Convo 3D (64)	3x3x3	8x29x57x64	Relu
MaxPool 2D	1x2x2	8x14x28x64	-
Convo 3D (128)	3x3x3	8x14x28x128	Relu
MaxPool 2D	1x2x2	8x7x14x128	-
Dense	64	-	-
Dense	32	-	-
Dense	5	-	Softmax

managed to increase the accuracy of vowel I, which had a 20% error when utilizing lip motion, reduced to only 0.47%, hinting that the vowel I challenge for detection with lip motion is resolved by analyzing the entirety of facial movement of the subject.

The facial movement recognition model also reduced the challenging vowel O which had an 18% error rate to 0.06% error rate, almost eliminating the error of this specific vowel. Vowel O indicated to have a sufficient distinction that can be found when the entirety of the subject's facial movement is analyzed, but this distinction is somewhat lost when it is focused on the lips of the subject.

However, both pieces of research are consistent in regards to vowels A and E, where prediction becomes a challenge high level of inaccuracy in both when compared to other vowels. It was a 17% error when lip motion is utilized, and then it provided approximately 5.4% error utilizing this research's model. The issue indicated that there is a need for an approach that allowed better differentiation between vowels A and E.

An important addition that this research provided is the capability of detecting a silent state, which allowed the detection of whether the speaker tried to speak a vowel or not. This expanded the model to not only differentiate the vowels but also to detect whether the subject is speaking a vowel or is silent.

IV. CONCLUSION

This research covered the implementation of vowel recognition by utilizing the facial movements of the speaking subject. A similar approach utilizing lip's motion is covered previously in research by Maxalmina [10], providing 84% highest accuracy. This research took the approach of utilizing facial movement with a CNN-BiLSTM model yielding a validation accuracy of 98.66%.

Accuracy improvement was also consistently followed for all vowels, yielding an error rate below 6%. The most notable improvement is noted for vowel I, O, and U with an error

TABLE IX
ERROR RATE COMPARISON BETWEEN LIP MOTION [10] AND THIS RESEARCH

Vowel	Maxalmina Lip Motion	Facial Movement Recognition
A	17%	5.47%
E	17%	5.38%
I	20%	0.47%
O	18%	0.06%
U	9%	0.71%

rate below 1%, while vowels A and E showed an error rate of around 5.4%. This result indicated a huge potential for facial movement recognition utilization in recognizing Bahasa Indonesia vowels in future research.

ACKNOWLEDGMENT

The author would like to thank Institut Teknologi Sepuluh Nopember, which had provided access to a Supercomputer needed to train the model efficiently.

REFERENCES

[1] G. Vigliocco, P. Perniss, and D. Vinson, “Language as a multimodal phenomenon: implications for language learning, processing, and evolution,” Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 369, no. 1651. The Royal Society, p. 20130292, Sep. 19, 2014. doi: 10.1098/rstb.2013.0292.

[2] R. Sultana and R. Palit, ”A survey on Bengali speech-to-text recognition techniques,” 2014 9th International Forum on Strategic Technology (IFOST), Cox’s Bazar, Bangladesh, 2014, pp. 26-29

[3] I. Papadimitriou, A. Vafeiadis, A. Lalas, K. Votis, and D. Tzovaras, ‘Audio-Based Event Detection at Different SNR Settings Using Two-Dimensional Spectrogram Magnitude Representations’, Electronics, 2020.

[4] S. Prom-on and M. Onsri, ”Effects of Facial Movements to Expressive Speech Productions: A Computational Study,” 2019 IEEE 2nd International Conference on Knowledge Innovation and Invention (ICKII), Seoul, Korea (South), 2019, pp. 481-484.

[5] Z. Lu and L. Czap, ”Modelling the tongue movement of Chinese Shaanxi Xi’an dialect speech,” 2018 19th International Carpathian Control Conference (ICCC), Szilvasvarad, Hungary, 2018, pp. 98-103, doi: 10.1109/CarpathianCC.2018.8399609.

[6] K. Kumatani and R. Stiefelhagen, ”State Synchronous Modeling on Phone Boundary for Audio Visual Speech Recognition and Application to Muti-View Face Images,” 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP ’07, Honolulu, HI, USA, 2007, pp. IV-417-IV-420.

[7] S. Isobe et al., ‘GAMVA: A Japanese Audio-Visual Multi-Angle Speech Corpus’, 2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), pp. 134–139, 2021.

[8] N. K. Mudaliar, K. Hegde, A. Ramesh, and V. Patil, ”Visual Speech Recognition: A Deep Learning Approach,” 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2020, pp. 1218-1221, doi: 10.1109/ICCES48766.2020.9137926.

[9] T. Tasaka and N. Hamada, ”Speaker dependent visual word recognition by using sequential mouth shape codes,” 2012 International Symposium on Intelligent Signal Processing and Communications Systems, Tamsui, Taiwan, 2012, pp. 96-101, doi: 10.1109/ISPACS.2012.6473460.

[10] Maxalmina, S. Kahfi, K. N. Ramadhani, and A. Arifianto, ”Lip Motion Recognition for Indonesian Vowel Phonemes Using 3D Convolutional Neural Networks,” 2020 3rd International Conference on Computer and Informatics Engineering (IC2IE), Yogyakarta, Indonesia, 2020, pp. 157-161, doi: 10.1109/IC2IE50715.2020.9274562.

[11] G. Inc., ”MediaPipe”, 2022. [Online]. Available: https://github.com/google/mediapipe.

[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, ‘Learning representations by back-propagating errors’, Nature, vol. 323, no. 6088, pp. 533–536, 1986.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, ”Gradient-based learning applied to document recognition”, Proceedings of the IEEE, vol. 86, no. 11, 1998, pp. 2278–2324.

[14] S. Hochreiter and J. Schmidhuber, ‘Long Short-term Memory’, Neural computation, vol. 9, pp. 1735–1780, 12 1997.

[15] S. Kanai, Y. Fujiwara, Y. Yamanaka, and S. Adachi, ‘Sigsoftmax: Reanalysis of the Softmax Bottleneck’, arXiv [stat.ML]. 2018.

[16] J. D. O’Connor, and J. L. M. Trim, “Vowel, Consonant, and Syllable—A Phonological Definition,” vol. 9, no. 2. Informa UK Limited, pp. 103–122, Aug. 1953.

[17] Kementrian Pendidikan dan Kebudayaan Indonesia, ”Huruf Vokal - EYD V”, 2023. [Online]. Available: https://ejaan.kemdikbud.go.id/eyd/penggunaan-huruf/huruf-vokal/.

[18] Kementrian Pendidikan dan Kebudayaan Indonesia, ”Kata Pengantar - EYD V”, 2023. [Online]. Available: https://ejaan.kemdikbud.go.id/eyd/.

[19] A. L. Maas, A. Y. Hannun, and A. Y. Ng, ”Rectifier Nonlinearities Improve Neural Network Acoustic Models”, 2013.

[20] B. Xu, N. Wang, T. Chen, and M. Li, “Empirical Evaluation of Rectified Activations in Convolutional Network,” May 2015.

[21] D. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” International Conference on Learning Representations, Dec. 2014.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” Journal of Machine Learning Research, vol. 15, no. 56, pp. 1929–1958, 2014.