Voice Controlled Wheelchair for Disabled Patients based on CNN and LSTM

Sutikno
Dept. of Electrical Engineering
University of Jember
Jember, Indonesia
email: sutikno@unmuhjember.ac.id

Khairul Anam
Dept. of Electrical Engineering
University of Jember
Jember, Indonesia
email: khairul@unej.ac.id

Azmi Saleh
Dept. of Electrical Engineering
University of Jember
Jember, Indonesia
email: azmi2009@gmail.com

***Abstract*—Disabled patients with reduced mobility due to health problems such as disability, injury, paralysis, or other factors will experience difficulty in movement. They need tools that can help them, in which the most widely used is a wheelchair. The main objective of this research is to control wheelchair motion with voice commands. There are five commands for wheelchair control: forward, backward, right, left, and stop. Voice data is obtained from recording several subjects using Sound Recorder Pro and Sox Sound Exchange. The voice commands for wheelchair navigation were identified using Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) combination embedded in Raspberry Pi 3. Voice data is first converted to spectrogram images before being fed into CNN using Mel-Frequency Cepstrum Coefficients (MFCC). This system can be controlled by simple voice commands given by the user. This method is proven to be useful in speech recognition with an accuracy level using CNN-LSTM above 97.80 %. Preliminary experimental results indicate that voice commands in wheelchairs using the CNN-LSTM can be recognized well.**

Keywords—disability, wheelchair, MFCC, CNN, LSTM

I. INTRODUCTION

Persons with disabilities have many inconveniences in their lives. The disability that they suffer from can be physical, mental, intellectual, or even sensory. These difficulties can limit them to fully and effectively participate in society [1]. According to a survey conducted by the World Health Organization (WHO), it estimated that more than 1 billion people in the world have a disability in some form of disability [2]. Meanwhile, the 2010 PUSDATIN data from the Ministry of Social Affairs in Indonesia, was 11,580,117 in which 3,010,380 of them were physically impaired [3].

In the biomedical sector, wheelchairs are a vital device because of the latest advancements in the industrial population. Demand for patients with particular limitations is increasing because intelligent wheelchairs will play an essential role in the welfare society in the future. The use of intelligent wheelchairs inspires machines' view as partners and not as instruments [4]. In previous studies, many articles discussed introducing voice commands for wheelchair navigation in prototype form, using Arduino, ANN, Fuzzy, and other intelligence microcontrollers [5] [6]. Nevertheless, only a few use comparisons between the Convolutional Neural Network (CNN) and Long short-term memory (LSTM) methods for speech recognition applied to wheelchairs. Besides, the prior study's main objective of this wheelchair system is the addition of electric wheelchairs using sound conversion technology to drive electric wheelchairs using the CNN-LSTM network embedded in Raspberry Pi 3.

In the future, this smart wheelchair will have a significant impact on the disabled. As a result, electric wheelchairs that use speech identification technology can help to direct wheelchairs and plan movements in a better way to facilitate patient life. The speech recognition system is used as a user interface in operating the system in the system presented. Patients who are categorically unable to walk and must use a wheelchair for their daily work can control the wheelchair's movement only with their voice commands. When using this wheelchair, people will be able to move more independently. The wheelchair control uses speech recognition to trigger and control all movements [7]. The main emphasis is undoubtedly to concentrate on improving the wheelchair by incorporating several new features as its advantages [8].

II. PROPOSED METHOD

In general, the block diagram is a general description of the system to be built. It is necessary to make a block diagram/process flow that can help the readers to understand the research's flow and process quickly. Fig. 1 indicates the process flow of this study.

A. Preparation of the dataset

The sound recording process was carried out to obtain a dataset. The recording used a Sound Recorder Pro application and Sound Meter as a tool to detect the ambient noise level. Some parameters and variables were recorded, as shown in table 1.

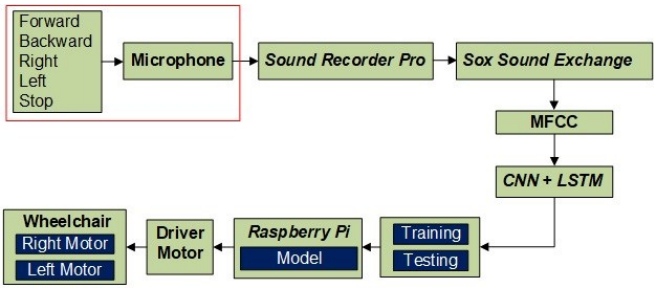

Fig. 1. The proposed block diagram

978-1-7281-9526-1/20/$31.00 ©2020 IEEE

TABLE I. VOICE RECORDING PARAMETERS

Variabel	*Information*
Data type	Sound (forward, backward, left, right, stop).
Sample rate, bitrate	16.000 Hz, 128 Kbit/s
The duration of each word	1 – 2 second.
Storage Format	*.wav (PCM)
Channels, sources	Mono/stereo/ microphone
Number of sound samples	1000 / word
Recording time	Adjusted
Recording location	Space is closed and open

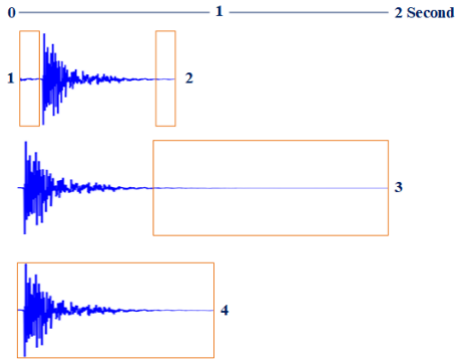

Fig. 2. Illustration of the trim and the pad process

In the table above, the recorded voice commands are forward, backward, right, left, and stop using a microphone with a sample rate of 16.000 Hz, assisted by using a Sound Recorder Pro application recorded in .wav format. Each command was recorded as many as 1000. Meanwhile, the location recording was done in a closed and open space with a noise level of 33.0 - 83.8 decibels. After the data was recorded, the process went to the preprocessing stage to get the same size using the notepad and Sox-Sound Exchange tools. This stage aims to make the data uniform and easier to use as input to the system to be created. The command to do is trim forward 0.1 seconds, trim back 0.1 seconds, pad 2 seconds, and trim 1 second using Sox-Sound Exchange, as shown in Fig. 2. Trim is the process of cutting data from time 0 to time t, while pad adds silence to the sound until the sound reaches t. From this process, the voice data is ready to be a dataset.

B. Mel-Frequency Cepstrum Coefficients (MFCC) Method

The next step is visualized data. Since the data is a one-dimensional vector, it makes the visualization process easier. However, it is necessary to know that the speech recognition process rarely uses raw data. Instead, it usually uses an approach of converting from an audio file into a spectrogram. The methods often used to analyze audio/voice signals are Mel-Frequency Cepstrum Coefficients (MFCC) and Short Time Fourier Transform (STFT) [9] [10]. In this study, the method employs MFCC to convert audio files into image spectrograms. The MFCC method is one of the domain methods and is most commonly used to perform feature extraction. MFCC is considered a frequency domain feature that is much more accurate than a time-domain feature for converting the voice signal into several parameters [11]. The reason why MFCC is chosen to be employed is the proof of its representation in signals. As a result, it becomes the most popular method in varied applications, which need voice preprocessing before being implemented. In its running process, MFCC works based on the difference in frequency of the human ear. Therefore, it can indicate the real human sound signals. Usually, the sampling frequency used is above 10.000 Hz to minimize the aliasing effect of the analog-digital conversion. The following Fig. 3 is how the MFCC process works.

The input form of voice data is in the .wav with a 16,000 Hz sample as test data and training data. Frame blocking is a condition where a signal is segmented into overlapping frames. Windowing aims to minimize signal friction of each frame starting from the beginning to the end. If the window is defined as w(n), $0 \leqslant n \leqslant N-1$, where N is the number of samples in each frame, the result of signal windowing is:

$$Y_1(n) = x_1(n)\, w(n),\ 0 \leq n \leq N\text{-}1 \tag{1}$$

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N\text{-}1}\right),\ 0 \leq n \leq N\text{-}1 \tag{2}$$

Where $w(n)$ is the resulting window frames, N is the number of sample values per frame, and n is the number of frames. The next step is a Fast Fourier Transform (FFT), which is a fast algorithm for implementing a discrete Fourier transform (DFT). FFT can change each sample N frame from the time to frequency domain, which is formulated bellow:

$$X_n = \sum_{k=0}^{N-1} x_k\ eps\left(-\frac{2\pi jkn}{N}\right),\ n = 0,1,2,\ldots\ldots N\text{-}1 \tag{3}$$

Xn is the frequency, xk is the sample value, N is the number of sample data, and j is the imaginary number. The subsequent step is the Mel-Frequency Wrapping stage. The measurement used to get voice signal with actual frequency f is usually in Hz. A subjective pitch is measured on a scale of "Mel". The scale of the Mel-frequency is a linear low frequency and a high logarithmic frequency, each consisting of under 1000 Hz and over 1000 Hz, respectively. This scale is defined as:

$$X_i = log_{10}\left(\sum_{k=0}^{N-1} |X(k)| H_i(k)\right),\ i=1,2,3\ldots\ldots,M \tag{4}$$

$H_i(k)$ is the value of the filter triangle I, $X(k)$ is the value of k data from the FFT process, M is the number of filters, and N is the number of data. The final step is to change the log mel spectrum into the time domain, which results in the Cepstrum Mel Frequency Coefficient (MFCC). Cepstrum is the opposite designation for a spectrum. Cepstrum is generally used to gather information from a voice signal spoken by humans. The log Mel spectrum as a final stage is changed to cepstrum using Discrete Cosine Transform (DCT). The following equations are used in the cosine transformation:

$$C_j = \sum_{i=0}^{M} X_i \cos\left(j(i-1)/2\frac{\pi}{M}\right) \tag{5}$$

j is 1,2,3,…. K (K is the number of coefficients intended and M is the number of filters). The results of the MFCC process for sound signals can be seen in Fig. 4.

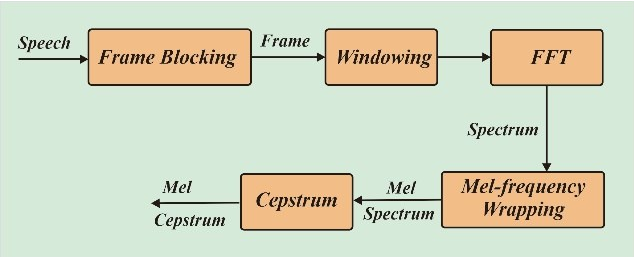

Fig. 3. Block diagram of MFCC process [12].

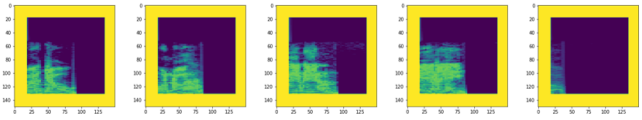

Fig. 4. Spectrogram of forwards, backwards, right, left, stop commands.

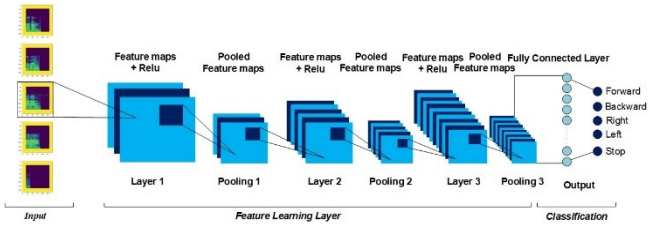

Fig. 5. Convolutional neural network architecture

C. Convolutional Neural Network (CNN)

The CNN method as a type of neural network is commonly used to proses any image data [13]. This method can detect any objects in the form of images, in which it is almost the same as ordinary neural networks. Its architecture comprises of neurons equipped with heavy, biased, and activated functions. Since CNN has more networks and it is widely applied to image data, it is usually categorized as Deep Neural Network. In its running process, there are two methods that have been categorized as feedforward and backpropagation. The way CNN works is similar to Multi-Layer Perceptron (MLP), where each neuron has only one dimension. However, each CNN neuron is shown in a two-dimensional form, making it different from MLP. The CNN architecture in this study is shown in Fig. 5. The CNN architecture that was built consisted of inputs, Feature Learning Layers, Fully Connected Layer, and Output as Classification.

Input for CNN is in the form of a spectrogram image. A matrix represents this image with a size of 20 x 11. Layer 1, layer 2, and layer 3 are the convolutional layer. Convolutional layer 1 consists of 32 filter layers where the convolution happens between receptive fields and kernels of 3 x 3 using the same padding and stride 1. This architecture is also applied to layer 2 and layer 3. All receptive fields will be in the partial overlap. All neurons will share the weight of the connection [14].

D. Long Short-Term memory (LSTM)

The LSTM is a particular type of Recurrent Neural Network (RNN), capable of studying long-term and short-term data. LSTM works through several cells, which are being parts of LSTM architecture. Each cell has three main components responsible for forgetting, remembering, and updating data [15]. Fig. 6 shows how the LSTM architecture is. Just like RNN, the LSTM also consists of modules with processing. What makes it different is that the LSTM modules are from themselves.

The LSTM decides any information needed to be removed from the C_{t-1} cell, taking advantage of a sigmoid gate called the "forget gate" f_t. This gate is capable of reading the values h_{t-1} and x_t, and producing a number between 0 and 1 for each element in C_{t-1}. A value of 1 means "really take care of this element", while 0 means "get rid of this element completely". The following is the forget gate and the gates element-wise sum formula:

$$f_t = \sigma (W_f . [h_{t-1} , x_t] + b_f \quad (6)$$

$$i_t = \sigma (W_i . [h_{t-1} , x_t] + b_i \quad (7)$$

$$\hat{C}_t = tanh (W_c . [h_{t-1} , x_t] + b_C \quad (8)$$

The last part of the LSTM is the output gate (neuron layer with the sigmoid activation function at the far right of the neuron layer line). This gate output does not contribute to the state of the cell, but it is this gate that differentiates the cell state and the actual output. The result of the LSTM can be seen in the following:

$$o_t = \sigma (W_o . [h_{t-1} , x_t] + b_0 \quad (9)$$

$$h_t = o_t . tanh (C_t) \quad (10)$$

The LSTM structure using voice commands is shown in Fig. 7.

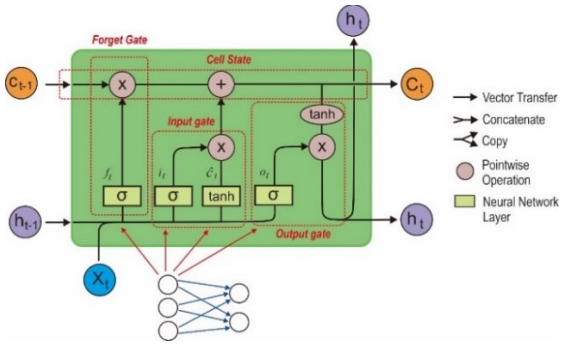

Fig. 6. The architecture of the LSTM.

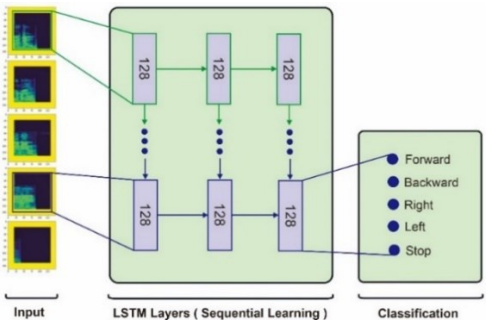

Fig. 7. LSTM architecture for speech recognition

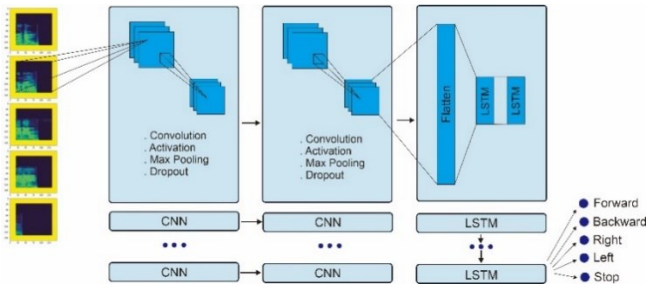

Fig. 8. Architecture of CNN-LSTM

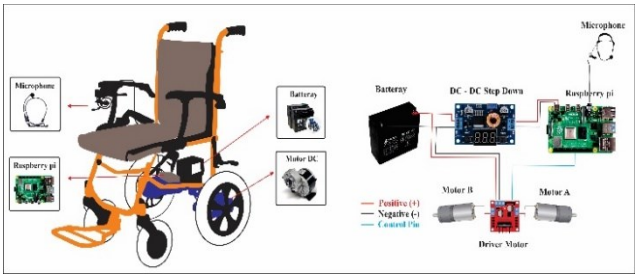

Fig. 9. Wheelchair hardware and electric design

E. CNN – LSTM Classifier

A way to analyze audio/voice signals is a 2-dimensional representation, where time-frequency analysis is commonly used to process sound signals. In this study, the voice signal is converted into 2D representation using Mel Frequency Cepstrum Coefficient (MFCC), which is done after preprocessing voice signals. It is then analyzed by CNN and LSTM [16]. Fig. 8 shows a detailed schematic of our proposed process.

F. Application for Wheelchairs

The CNN is applied in a wheelchair after the offline training and testing are completed. It aims to perform better accuracy. Then, the next process is to save the CNN model and then embed it into the raspberry pi. There is only a raspberry pi model produced from the CNN model built. As a result, it is not heavy, and the dataset does not need to be embedded in raspberries. Data on the raspberry pi is digital data for moving wheelchairs according to user orders, where data is sent to the motor driver to drive the motor in a wheelchair. The next step is to design a wheelchair and its supporting devices. Fig. 9 shows the hardware and electrical design of a wheelchair.

III. Results and Discussion

This study discusses the control of a wheelchair using CNN-LSTM as a new control method through its combination offered. There are five types of wheelchair commands: forward, backward, right, left, and stop. Each voice command is taken as many as 1000 data. As a result, total data of 5000 data were acquired and then divided into training data of 4000 and testing data of 1000. For determining the proposed method's performance, two tests of CNN and CNN-LSTM were carried out. Initial testing used CNN with input shape (180.1) and 50 epochs.

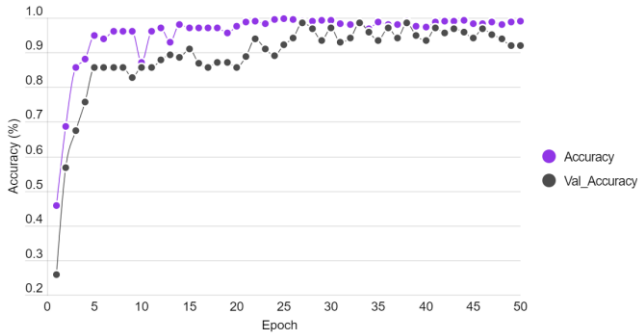

Fig. 10. Accuracy using the CNN method

Fig. 10 shows the accuracy level using CNN. The training stage yielded an accuracy of 98.95%, while the testing stage obtained an accuracy of 92.00%. Each processed voice command had a different level of accuracy, as shown in Table II. It occurred because the signal pattern recognition process on each label was different.

The results will be displayed in the form of a confusion matrix to describe the classification model's performance. The following is Fig. 11, which describes the confusion matrix of using CNN. The next test is the combination of CNN-LSTM. By using the same input shape and epoch, here are the test results. This test aims to obtain a better level of accuracy for speech recognition. Fig. 12 shows a graph of the accuracy level using CNN-LSTM.

TABLE II. The Voice Command Recognition Process Using The CNN Method

Commands	*Precision*	*Recall*	*F1-Score*	*Support*
Forward	1.00	1.00	1.00	186
Backward	1.00	1.00	1.00	223
Right	1.00	0.62	0.76	210
Left	1.00	1.00	1.00	194
Stop	0.70	1.00	0.82	187
Accuracy	-	-	0.92	1000
Macro Average	0.94	0.92	0.92	1000
Weighted Average	0.94	0.92	0.92	1000

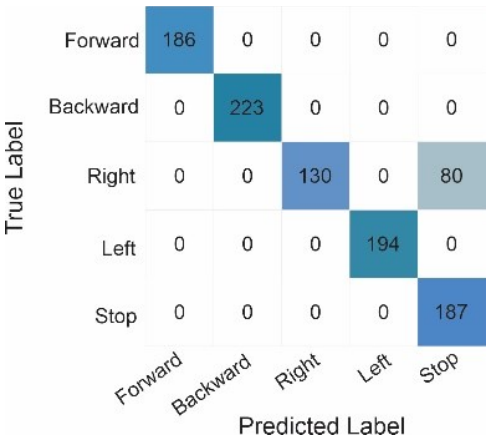

Fig. 11. Confusion matrix voice command recognition with CNN

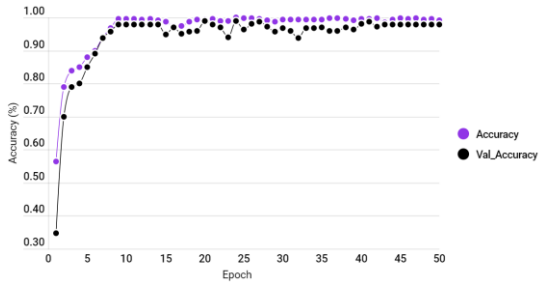

Fig. 12. The level of accuracy uses the CNN-LSTM method

TABLE III. THE VOICE COMMAND RECOGNITION PROCESS USES THE CNN METHOD

Commands	*Precision*	*Recall*	*F1-Score*	*Support*
Forward	1.00	1.00	1.00	186
Backward	1.00	1.00	1.00	223
Right	1.00	0.90	0.94	210
Left	1.00	1.00	1.00	194
Stop	0.89	1.00	0.94	187
Accuracy	-	-	0.98	1000
Macro Average	0.98	0.98	0.98	1000
Weighted Average	0.98	0.98	0.98	1000

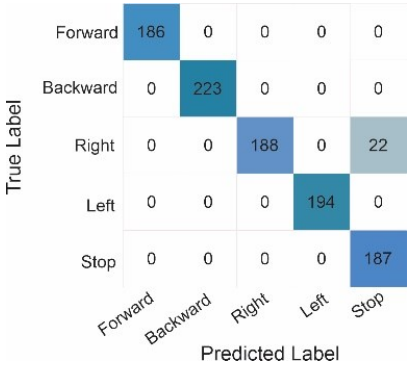

Fig. 13. Confusion matrix voice command recognition with CNN-LSTM

The comparison accuracy of both methods is not too significant. The CNN obtained an accuracy of 98.95% with validation accuracy of 92%. However, the combination methods of CNN-LSTM resulted in slightly higher accuracy of 99.15% with validation accuracy of 97.80%. The results of the test process on each voice command and its accuracy are shown in Table III. With a good level of accuracy above, voice commands using CNN-LSTM applied to wheelchairs were intended to get high performance. Nevertheless, there were still voice commands that experienced errors in label recognition, as shown in Fig. 13 below. According to Fig. 13 above, there are no changes that are much different either by using CNN or CNN-LSTM for voice commands of forward, backward, left, right, and stop. Here the error is the voice command Right. If using CNN, this command reads 88 as a stop command. Meanwhile, the CNN-LSTM was less than 22.

CONCLUSION

The application of voice commands to an electric wheelchair using the CNN-LSTM method has a better accuracy level, instead of using CNN. This method can recognize voice commands given by users with an accuracy rate of 97.80%. This result is influenced by the model of CNN-LSTM built. The accuracy for each method is different because of the structure and parameters in the model, such as learning rate, activation, and dropout. Eventually, this sound-based wheelchair can help the disabled to move more easily without wasting much energy.

ACKNOWLEDGMENT

We want to thank the lecturers and friends, especially at the master's program of electrical engineering, University of Jember, who participated in writing this article to complete it on time.

REFERENCES

[1] M. B. Santoso and N. C. Apsari, "Pergeseran paradigma dalam disabilitas," vol. 1, no. 2, pp. 166–176, 2017, doi: 10.24198/intermestic.v1n2.6.

[2] C. Joseph, S. Aswin, and J. Sanjeev Prasad, "Voice and gesture controlled wheelchair," Proc. 3rd Int. Conf. Comput. Methodol. Commun. ICCMC 2019, no. Iccmc, pp. 29–34, 2019, doi: 10.1109/ICCMC.2019.8819662.

[3] P. Arie, "Inklusi Penyandang Disabilitas di Indonesia," J. Refleks. Huk., vol. 1, pp. 1–4, 2017.

[4] K. P. Tiwari and K. K. Dewangan, "Voice Controlled Autonomous Wheelchair," Int. J. Sci. Res., no. April, pp. 10–11, 2015.

[5] M. H. Jabardi, "Voice Controlled Smart Electric-Powered Wheelchair Based on Artificial Neural Network Network," Int. J. Adv. Res. Comput. Sci., vol. 8, no. 5, pp. 31–37, 2017.

[6] M. M. Abdulghani, K. M. Al-Aubidy, M. M. Ali, and Q. J. Hamarsheh, "Wheelchair Neuro Fuzzy Control and Tracking System Based on Voice Recognition," Sensors (Basel)., vol. 20, no. 10, 2020, doi: 10.3390/s20102872.

[7] P. C A, N. Vijay, and P. Bisen, "Voice Controlled Wheelchair for Physically Disabled People," Int. J. Res. Appl. Sci. Eng. Technol., vol. 6, no. 5, pp. 2375–2380, 2018, doi: 10.22214/ijraset.2018.5389.

[8] C. N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," Artif. Intell. Rev., vol. 43, no. 2, pp. 155–177, 2012, doi: 10.1007/s10462-012-9368-5.

[9] L. Beth, "Mel-Frequency Cepstral Coefficients For Music Modeling," Cambridge Res. Lab., no. c, pp. 2–6.

[10] A. Elbir, H. O. Ilhan, G. Serbes, and N. Aydin, "Short Time Fourier Transform based music genre classification," 2018 Electr. Electron. Comput. Sci. Biomed. Eng. Meet. EBBT 2018, pp. 1–4, 2018, doi: 10.1109/EBBT.2018.8391437.

[11] N. Dave, "Feature Extraction Methods LPC , PLP and MFCC In Speech Recognition," Int. J. Adv. Res. Eng. Technol., vol. 1, no. Vi, pp. 1–5, 2013.

[12] H. S. Kumbhar and S. U. Bhandari, "Speech emotion recognition using MFCC features and LSTM network," Proc. - 2019 5th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2019, pp. 2019–2021, 2019, doi: 10.1109/ICCUBEA47591.2019.9129067.

[13] Prabhu, "Understanding of Convolutional Neural Network (CNN) — Deep Learning," 2018. [Online]. Available: https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148.

[14] Suyanto, K. Nur Ramadhani, and S. Mandala, DEEP LEARNING : Modernisasi Machine Learning Untuk Big Data. Bandung: Informatika Bandung, 2019.

[15] M. Hajiaghayi and E. Vahedi, "Code failure prediction and pattern extraction using LSTM networks," Proc. - 5th IEEE Int. Conf. Big Data Serv. Appl. BigDataService 2019, Work. Big Data Water Resour. Environ. Hydraul. Eng. Work. Medical, Heal. Using Big Data Technol., no. c, pp. 55–62, 2019, doi: 10.1109/BigDataService.2019.00014.

[16] W. Lim, D. Jang, and T. Lee, "Speech Emotion Recognition using Convolutional and Recurrent Neural Networks," 2016 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf., 2017, doi: 10.1109/APSIPA.2016.7820699.