Autonomous Voice Recognition Wheelchair Control System

Mokhles M. Abdulghani
Jackson State University
Jackson, Mississippi
mokhles.m.abdulghani@students.jsums.edu

Wilbur L. Walters
Jackson State University
Jackson, Mississippi
wilbur.l.walters@jsums.edu

Khalid H. Abed
Jackson State University
Jackson, Mississippi
khalid.h.abed@jsums.edu

Abstract*—Autonomous wheelchairs are essential for improving the mobility of individuals with disabilities or physical challenges. Advances in computer and wireless communication technologies have paved the way for the design of smart wheelchairs to match the needs of disabled people. This research paper introduces the design and implementation of an electric wheelchair controller using voice recognition. This model is based on a voice recognition algorithm for classifying the essential commands to drive the wheelchair. A pre-designed adaptive neuro-fuzzy controller has been used to generate the required control signals for activating the motors of the wheelchair. This controller alters data gathered from obstacle avoidance sensors and a voice recognition classifier. Deep learning algorithms have made many technological developments and influences modern daily lives. After testing our developed deep learning algorithm, the overall classification accuracy for distinguishing between eight voice commands was 98 %.

Keywords: *Voice Recognition, VR, Deep Learning, DL, EEG, Brain-Computer Interface, BCI, Brain-Machine Interface, RNN, BMI, LSTM, STFT, Autonomous Wheelchairs.*

I. INTRODUCTION

Speech could be utilized as a user interface to cooperate with machines, such as wheelchairs and robots. It has been made viable to have improved systems with capabilities for real-time conversations. However, this is facing a lot of problems, which are due to the difference in speaker's speech caused by age, gender, speed of speech, different accent, environmental noise, and more [1]. Voice recognition is the capability of a political machine or software to receive and translate dictation or to understand and bring outspoken commands. The first voice recognition invention was introduced in 1990 by Dragon. As presented in the literature [1 – 3], the first voice recognition system that could recognize continuous speech was introduced by IBM in 1996. Throughout the past twenty years, there has been exponential development in voice-controlled products, especially after the release of smartphones, where more complex voice recognition software products have been built. Using voice will not serve all individuals who need wheelchair assistance, as for people with high levels of disability using brain thoughts might be the best solution [4 – 6]. Understanding the extent of this problem has motivated us to research far beyond just engineering and to dive into biomedical and psychology to understand the extent of how brain neurons work and to understand the reasons that cause a person's specific neurons to fire and then manipulate the signal.

To overcome the need of using a joystick or any other input procedure that requires moving muscles (specifically for those experiencing a high level of disability), this paper introduces a voice-based wheelchair control system for disabled or physically challenged people. Voice recognition systems are classified into two types, namely speaker-dependent and speaker-independent. The speaker-dependent system is based on training the person, who will be using the system, while the speaker-independent system is trained to respond to a word regardless of who speaks. The first type demonstrates high accuracy for word recognition; consequently, it is recommended for a voice-controlled wheelchair. Speech is one of the most valuable methods of communication among humans. By employing a microphone sensor, speech can be used to interact with a computer and serve as a potential method for Human-Computer Interactions (HCI) [7].

II. RELATED WORK

A voice recognition unit is required to provide a communication channel between the computer and human voice. This interface is mainly based on the feature extraction of the desired sound wave signal. A typical voice recognition system consists of a data acquisition system, pre-emphasis of the acquired signals, feature extraction process, classification of the features, post-processing of the classifier output, and finally the control interface and device controller. To defeat the lack of accuracy in recognizing and classifying one's speech, numerous researchers have used the Convolutional Neural Network (CNN) method for voice recognition skills [8 – 9]. In [10], researchers examined 2D feature spaces for voice recognition built on CNN. The results demonstrated that the maximum rate of a word recognition was achieved using spectral analysis. Furthermore, the Mel scale (a scale of pitches judged by listeners to be equal in distance one from another) and spectral linear cepstral are outperformed by cepstral feature spaces. Huang, et al. [11] suggested a technique to analyze CNN for speech recognition by visualizing the local filters learned in the convolutional layer to detect routine learning. This method has improved in

2769-5654/22/$31.00 ©2022 IEEE
DOI 10.1109/CSCI58124.2022.00043

distinguishing four domains of CNNs. These domains are faraway speech recognition, low footprint, noise robustness, channel-mismatched, and training–test conditions.

The steering of smart wheelchairs using voice recognition skills with CNN has drawn many researchers [12]. Sutikno, et al. [8] presented a voice control system for wheelchairs using Long Short-Term Memory (LSTM) combined with CNN. The accuracy level of this method was about 97 %. Further research was performed by Ali, et al. [13], who constructed an algorithm for smart wheelchairs using CNN to assist people with disabilities in spotting bus doors. The method was executed based on accurate localization data and used a high-performance microprocessor or CPU for fast detection. Nevertheless, the use of CNN in smartphones is yet under progress due to linked complicated computations to achieve high-accuracy calculations [14]. A convolutional neural network algorithm has been used to distinguish between pre-recorded eight voice commands that have been downloaded from a Google audio library via MATLAB 2023a [15].

III. THE PROPOSED SYSTEM

The proposed autonomous wheelchair system consists of four main elements, an electric wheelchair, a voice recognition unit, real-time control unit, and an autonomous obstacle avoidance unit, as illustrated in Figure 1. The recorded voice is delivered to the voice recognition unit, which will verify the required action, based on his/her voice. A secondary control unit (microcontroller) will communicate serially with the intelligent voice recognition unit. The navigation and steering of the wheelchair have been controlled using a predesigned Adaptive Neuro-Fuzzy Inference System (ANFIS) uploaded on the main control unit.

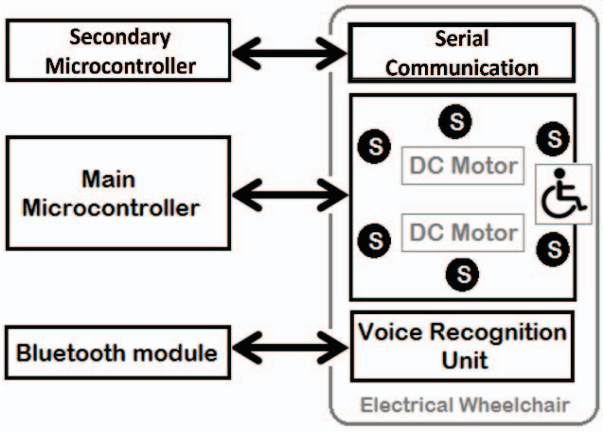

Figure 1: The proposed wheelchair steering system.

The voice recognition unit has been constructed using deep learning as follows:

A. Preparing data

The data consist of a 2500 voice audio clip for each of the eight commands used to steer the wheelchair, which is: (on, off, go, left, right, up, down, and stop). (On and off) to switch between the drive and parking status, (go and stop) for driving forward and stopping the wheelchair, (left and right) to change direction, and (up and down) have been added as extra commands, which can be used to increase or decrease the wheelchair speed, or can be used as a safety option to control the acceleration applied on the wheels during up-hill and down-hill driving to prevent wheels' slippage. The data has been split into training and testing data (80 % for training and 20 % for testing). The required voice commands data have been downloaded from MATLAB 2023a audio library. All the data have been labeled, and all the other words that are not the required commands have been labeled as" unknown". Labeling words that are non-commands as "unknown" creates a group of words that approximates the delivery of all words other than the commands. The networks employ this group to learn the distinction between commands and all other words. To reduce the class disparity between the "known" and the "unknown" words, and accelerate processing, only a fraction of the unknown words has been included in the training set. Background noise has been added in a separate step later to enhance the model accuracy in the real-time execution.

B. Convert voice commands to auditory spectrogram

All the training and testing voice commands, including the "Unknown", have been converted to an auditory-based spectrogram, which is the visual representation of the audio (picture of audio) for more efficient training performance of the convolutional neural network. This has been done by splitting the audio into overlapping windows of 0.02 seconds in length and 16000 Hz frequency, performing the Short Time Fourier Transformation (STFT) on each window, and converting the resulting window to decibels. This provides us with a powerful image of the sound's shape. Finally, sending back these windows into the length of the initial voice command and presenting the output in his visual shape as shown in Figure 2. Some files in the data set are less than 1-second long, and others are more than 3 seconds, but the required input should hold a consistent size where all the data have the same length to be trained in its picture form by a CNN. Therefore, zero-padding has been applied to the beginning and end of each audio signal so that it is of the same length and ready to be an input for the CNN layers.

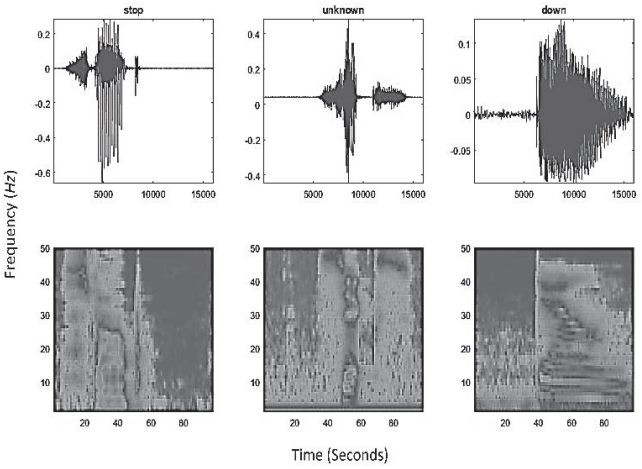

Figure 2: Spectrogram of the audio commands.

C. Training The CNN

The visual shape of the voice commands is now ready to be used to train a multi-layer convolutional neural network. The CNN is mainly constructed to deal with image distinguishing, which is the case we deal with (picture shapes of the eight-voice commands). The convolutional neural network model has been designed as six convolutional layers, each followed by batch normalization, Relu activation, and max-pooling layers. The network has ended with a drop-off layer that will randomly shut off 20 % of the training parameters (dropout ratio 0.2) to prevent overfitting in training data during the training process.

Another technique has been added to overcome the overfitting issue and for a smoother training process, an L_2 *Regularization.* This technique works on improving the calculation for the weights to reduce the loss function E(θ) and to reduce overfitting. The regularization term is also called weight decay. The loss function with the regularization term takes the form:

$$E_R(\theta) = E(\theta) + \lambda\Omega(w) \dots (1)$$

$$\Omega(w) = \frac{1}{2}w^T w \quad \dots (2)$$

where w is the weight vector, λ is the regularization coefficient (has been set to 0.0001), and the regularization function is $\Omega(w)$. Finally, the network has ended with Fully Connected, SoftMax, and Classification output layers with the eight class labels (equivalent to the required number of outputs) as illustrated in Figure 3. The accuracy that has been achieved in the training process for eight voice commands (on, off, go, left, right, up, down, stop) is shown in Figure 4.

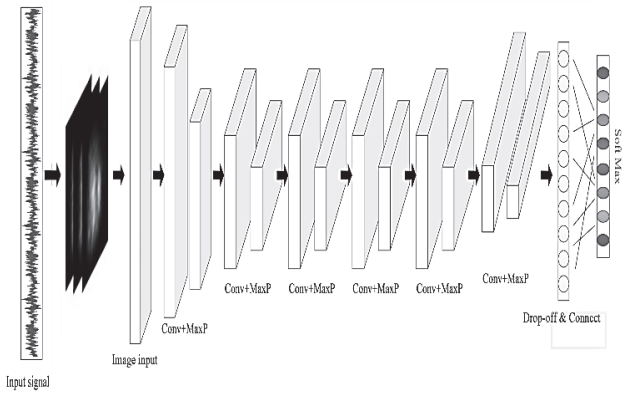

Figure 3: The structure of the CNN model.

IV. RESULTS AND ANALYSIS

Overall testing accuracy of 98 % (1.9976 % of validation error) has been achieved when testing the resulting convolutional neural network with the remaining 20 % dataset. These results have been obtained with the help of the adaptive moment estimation (Adam) algorithm. Adam algorithm has been used as an optimization algorithm to tune the hyper-parameters of the convolutional neural network model. After training the CNN model on 80 % of the dataset with 28 max epochs, 0.00019 learning rate, and a mini-batch size of 136, the 98 % accuracy has been archived. Figure 4 shows the accuracy in the training stage. Table 1 shows the accuracy obtained from the testing data for the designed convolutional neural network model to classify the eight voice commands.

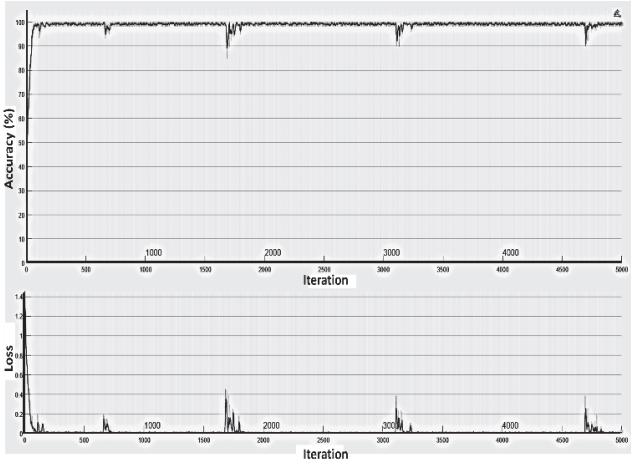

Figure 4: The training accuracy of the Convolutional neural network.

Table 1. The predicted outputs accuracy of the convolutional neural network.

Voice Command	Predicted Output Accuracy
on	96,4 %
off	98,3 %
go	98,2 %
left	98,1 %
right	95,8 %
up	99,5 %
down	99,4 %
stop	98,4 %

V. CONCLUSION AND FUTURE WORK

Employing the concept of deep learning and machine learning has shown acceptable to good accuracy (0.98) in distinguishing between eight commands (on, off, go, left, right, up, down, stop) represented by one's voice. Although this research showed a promising method for controlling wheelchairs using the human voice, more efforts are still needed to come up with a more reliable algorithm that can recognize the user's voice signature. Without recognizing the user's voice signature, using the resulting algorithm in this research can be dangerous as anyone close to the user can control the wheelchair by saying any one of the eight trained voice commands.

Further work will be considered to improve the accuracy and design of the next stage voice recognition unit, which can recognize the user's voice signature so no one can control the wheelchair but the user himself. The desired goal will be almost 100 % accuracies in the convolutional neural network algorithm. Moreover, a real-time test will be performed to come out with a real system.

REFERENCES

[1] Mazo, M.; Rodriguez, F.J.; Lazaro, J.L.; Urena, J.; Garcia, J.C.; Santiso, E.; Revenga, P.; Garcia, J.J. Wheelchair for Physically Disabled People with Voice Ultrasonic and Infrared Sensor Control. Auton. Robot. 1995, 2, 203–224.

[2] Geuaert, W.; Tsenav, G.; Mlad, V. Neural Network used for Speech Recognition. J. Autom. Control 2010, 20, 1–7.

[3] Rani, P.; Kakkar, S.; Rani, S. Speech Recognition using Neural Network. In Proceedings of the International Conference on Advancement in Engineering and Technology, ICAET 2015, Incheon, South Korea 11–13 December 2015; pp. 11–14.

[4] K. M. Al-Aubidy and M. M. Abdulghani, "Wheelchair Neuro Fuzzy Control Using Brain Computer Interface", *12th International Conference on Developments in eSystems Engineering (DeSE)*, pp. 640-645, 2019.

[5] H. Raad, F. Fargo, O. Franza, "Autonomic Architectural Framework for Internet of Brain Controlled Things (IoBCT)". ISNCC 2021.

[6] Al-Aubidy, Kasim M., and Mokhles M. Abdulghani. "Towards Intelligent Control of Electric Wheelchairs for Physically Challenged People." *Advanced Systems for Biomedical Applications*: 225.

[7] Bakouri, M.; Alsehaimi, M.; Ismail, H.F.; Alshareef, K.; Ganoun, A.; Alqahtani, A.; Alharbi, Y. Steering a Robotic Wheelchair Based on Voice Recognition System Using Convolutional Neural Networks. Electronics 2022, 11, 168.

[8] Anam, K.; Saleh, A. Voice Controlled Wheelchair for Disabled Patients based on CNN and LSTM. In Proceedings of the 2020 4th International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Indonesia, 10–11 November 2020; pp. 1–5.

[9] Sharifuddin, M.S.I.; Nordin, S.; Ali, A.M. Comparison of CNNs and SVM for voice control wheelchair. IAES Int. J. Artif. Intell. 2020, 9, 387.

[10] Korvel, G.; Treigys, P.; Tamulevicus, G.; Bernataviciene, J.; Kostek, B. Analysis of 2d feature spaces for deep learning-based speech recognition. J. Audio Eng. Soc. 2018, 66, 1072–1081.

[11] Huang, J.T.; Li, J.; Gong, Y. An analysis of convolutional neural networks for speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 4989–4993.

[12] Lee, W.; Seong, J.J.; Ozlu, B.; Shim, B.S.; Marakhimov, A.; Lee, S. Biosignal Sensors and Deep Learning-Based Speech Recognition: A Review. Sensors 2021, 21, 1399.

[13] Ali, S.; Al Mamun, S.; Fukuda, H.; Lam, A.; Kobayashi, Y.; Kuno, Y. Smart robotic wheelchair for bus boarding using CNN combined with Hough transforms. In Proceedings of the international Conference on Intelligent Computing, Wuhan, China, 15–18 August 2018; pp. 163–172.

[14] Martinez-Alpiste, I.; Casaseca-de-la-Higuera, P.; Alcaraz-Calero, J.M.; Grecos, C.; Wang, Q. Smartphone-based object recognition with embedded machine learning intelligence for unmanned aerial vehicles. J. Field Robot. 2020, 37, 404–420.

[15] Warden P. "Speech Commands: A public dataset for single-word speech recognition". Available from: https://storage.googleapis.com/download.tensorflow.org/data/speech_commands_v0.01.tar.gz. 2017.