SCHOLARLY DEBUT
by Scholarly Review

Spring 2026

Voice-Controlled Vehicle: Determining the Best Machine Learning Models for Voice Recognition

By Dhruv Panchagatti

AUTHOR BIO

Dhruv Panchagatti is currently a senior at Northwood High School. He enjoys engineering and soccer with a passion for building ML projects that connect stats to the real world. He hopes to work in Formula One someday.

ABSTRACT

Technology that helps visually impaired people often use mobility canes to navigate. Assistive navigation technologies also often rely on voice-based control systems, which can be vulnerable to unintended activations in shared environments. This research develops a one- and two-stage machine learning voice model to determine what the most accurate one is using tests on a miniature vehicle. The models were trained on a google dataset as well as user voice recordings to determine model coefficients. The models take raw wav files as input and employ a convolutional neural network to classify spoken directional commands like “go”, “down”, “left”, and “right”. In the two-stage model, the audio input gets classified as user’s audio or not user’s audio and the command is ignored if it isn’t the intended user’s audio. If the command is identified as the intended user’s voice, then it goes to the second stage for command classification. This was just an addition to see how one stage and two stages compare. It was found that the two-stage recognition maintained greater accuracy in command recognition and better resistance to command hijacking from unintended users. These results suggest that multi-stage classification outperforms single-stage voice classification when used strategically.

Keywords: *Machine Learning, Voice Recognition, Assistive Technology, Audio, Data, Visually Impaired, Disabilities, Command Recognition*

SCHOLARLY DEBUT
by Scholarly Review

Spring 2026

INTRODUCTION

The modern world is moving towards better assistive technology for all types of disabilities (WHO, 2022). Visual, aural, and physical aids are all examples of assistive technology (Cook & Polgar, 2015). Real-world applications could be an increased number of wheelchair ramps around busy areas, tactile paving at crosswalks, and even just captions for videos (WHO, 2022). The goal of assistive technology is accessibility not complexity so any piece of innovation that lets these people get the most out of their lives is valuable (WHO, 2022). But, sometimes complexity is necessary and beneficial for the impaired. Things like robotic prosthetics and smart wheelchairs provide mobility and freedom that would be impossible otherwise (Cook & Polgar, 2015). The technology is constantly evolving and will continue to do so (WHO, 2022).

However, one piece of technology has remained largely unchanged and it is known as the mobility cane. Visually impaired people use it to navigate the outdoors using tactile input as a way to map out what is in front of them (Dakopoulos & Bourbakis, 2010). While there has been innovation for this product like smart canes, there must be other adaptable tools that move individually from the human (Bourbakis, 2008). Canes have limited reach and information gathering capabilities (Dakopoulos & Bourbakis, 2010). Not to mention, they are effectively extensions of the human arm meaning they can be dangerous if pointed in the wrong direction or situation (Dakopoulos & Bourbakis, 2010).

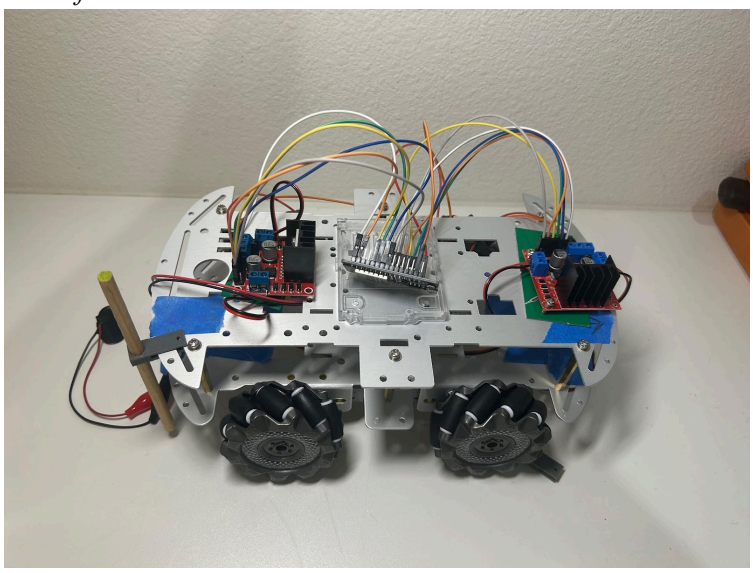
This is exactly where robot assistance and machine learning is capable of helping. Machine learning algorithms allow the robot to adapt and perform better over time allowing an infinite number of use cases and applications (Goodfellow, Bengio, & Courville, 2016). Most importantly, they are hands-free and separated from the user making them safer (Tapus, Mataric, & Scassellati, 2007). They can also be trained to listen to the intended user specifically and ignore other voices which increases safety (Tapus et al., 2007).

The mini vehicle that this research is studying uses simple commands that are interpreted by different machine learning voice models. Instead of measuring voice accuracy on an online app, it is more valuable to bring in real-world applications to account for some of the unknowns that come with assistive technology (WHO, 2022).

METHODOLOGY

In order to test out different voice models, the actual hardware for the vehicle is very important. For the chassis, two aluminum sheets were stacked on top of each other using small metal pillars allowing space for the batteries and power supply to be housed. This also allowed four motors to be attached to the underside of the chassis and ample space for the motor drivers and esp32 microcontroller to be placed on top. The esp32 was chosen specifically because it allows over-internet connection between itself and the web app. The machine learning models are loaded on to a simple web app which can be opened on a phone, laptop, etc. The built-in microphone on these devices will pick up the sound and use that as input into the chosen model.

Figure 1
View of robot vehicle



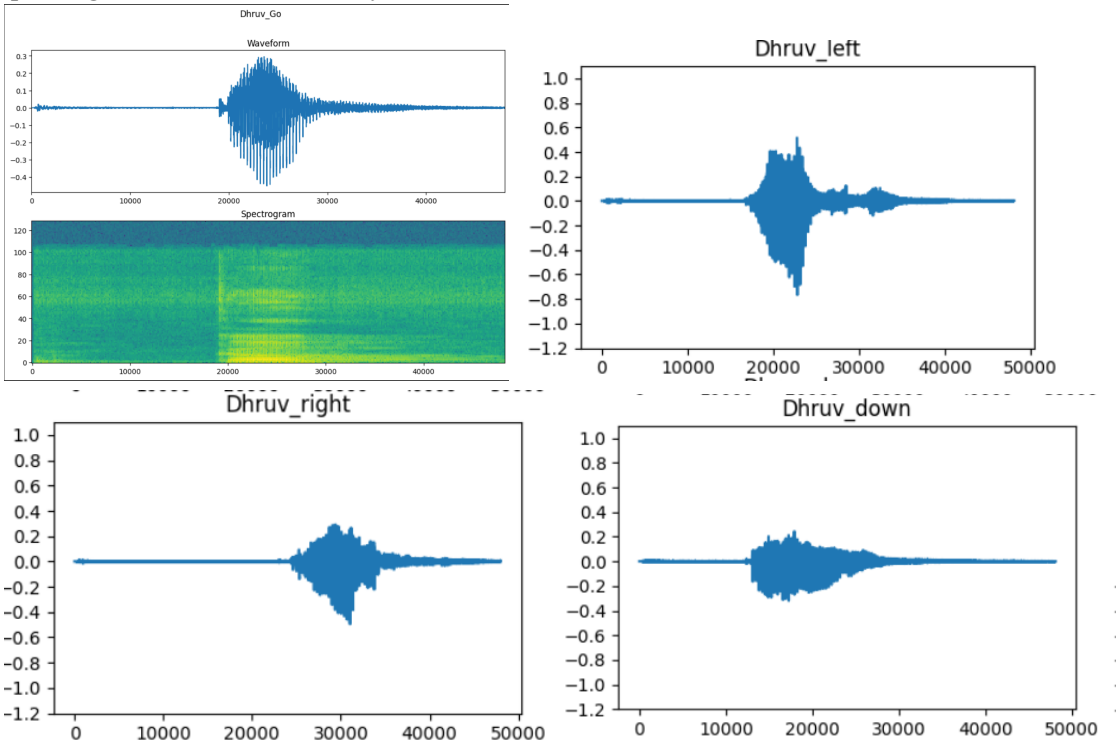
Note. Hidden in the middle layer are a breadboard and two battery packs. One to power the motor drivers and one to power the microcontroller. Four Mecanum (Omni) wheels allow full range of motion.

There are two machine learning models being evaluated: a one-stage model and a two-stage model. The one-stage model takes in the audio sample that the user speaks into the microphone and attempts to classify the command as either “go”, “down”, “right”, “left”, or background noise (go and down respectively mean forward and backward but were the only words available in the chosen dataset with a similar meaning to forward and backward). The model was trained on fifty samples (ten “go”s, ten “down”s, ten “right”s, ten “left”s, ten backgrounded noises) of intended user data and fifty samples (same spread) of audio samples from a kaggle dataset (Kurelli, 2021). These two datasets were merged into 100 samples which were split 80% into a training group and 20% into a validation group. The actual weights for the model were determined by the training group and then validated by the validation group based on how well it performed.

Specifically, the model is a Two Dimensional Convolutional Neural Network. The input is raw wav data which is converted into spectrograms.

Figure 2

Spectrograms and Wav Forms of Test Data



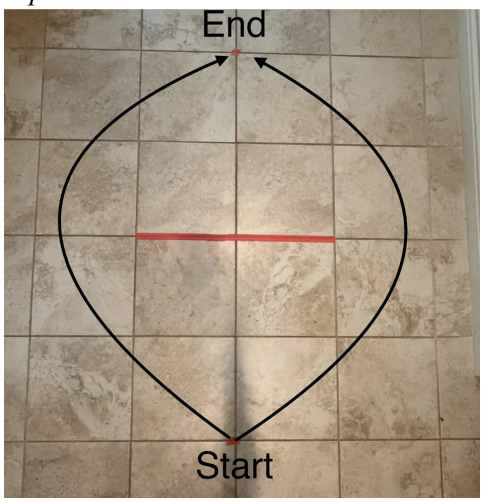
This process is then followed by preprocessing to ensure that all received audio is chopped up into one second audio chunks that can be better interpreted by the model.

The two-stage model essentially adds on to the functionality of the one-stage model. The only difference is that before the command classification, the first stage is to determine whether or not the audio matches the intended user's voice before classifying the command. Then, the second stage is the actual command classification to determine "go", "down", "left", and "right".

Once the coefficients and weights for both models were generated, a simple experiment was devised. Signs of effective models are accurate command recognition and resistance to external audio interference.

A vertical wooden dowel was placed on the robot to use as a marker for the position of the robot. Before every trial, the robot would be placed on the start point meaning the dowel would be placed directly over a marked circle on the starting piece of tape. The goal for the robot was to make it all the way to the finish tape without hitting the large red middle tape (left or right around it).

Figure 3
Experiment Track



Note. The arrows show possible paths for the vehicle to take around the red line.

The intended user speaks command after command guiding the robot from start to finish. But, between each intended user command, an external person spoke the same exact command again. These two factors test both user recognition and command accuracy within the same experiment. Each model had three trials on the track in order to make the data more rigorous.

RESULTS

Table 1
Experimental Statistics

Architecture	Commands Tested	Overall Error Rate	Hijack Rate
One-Stage CNN (Command Only)	**71**	**12.68%**	**14.08%**
Two-Stage CNN (Speaker + Command)	**97**	**8.25%**	**8.25%**

The two main statistics of importance are the overall error rate and the hijack rate of each model because those show both how accurate the command recognition and user distinction for each model is. The disparity in the number of commands tested simply comes from ultimately different paths being taken from start to finish requiring different numbers of commands. The overall error rate for the two-stage model is more than four percentage points lower and the hijack rate is almost six percentage points lower than the one-stage model. This shows that both command recognition accuracy and the resistance to command hijacking increases when a second stage of classification is added.

Table 2
One-Stage and Two-Stage Statistics

One Stage			
Trial	**Commands Tested**	**Hijacked Commands**	**Hijack Rate**
Trial 1	**25**	**4**	**16.00%**
Trial 2	**18**	**3**	**16.67%**
Trial 3	**28**	**3**	**10.71%**
Total	**71**	**10**	**14.08%**
Two Stage			
Trial	**Commands Tested**	**Hijacked Commands**	**Hijack Rate**
Trial 1	**39**	**6**	**15.38%**
Trial 2	**28**	**0**	**0.00%**
Trial 3	**30**	**2**	**6.67%**
Total	**97**	**8**	**8.25%**

When looking at the trial-level statistics, the two-stage model had zero hijacked commands in its second trial showing capability for flawless resistance to hijacking at times. In contrast, the one-stage model had pretty consistent hijack rates across the entire experiment.

DISCUSSION

The experiment conducted during the study and the data gathered from it shows that two-stage voice classification, where the two stages are user recognition followed by command recognition, is both more resistant to external hijacking and more accurate in command recognition.

There may be a few possible explanations for this. One is that having a two-stage recognition system allows the neural network to split up tasks into multiple stages. Having simpler tasks being separate allows the model to assign more accurate weights for the task at hand rather than generalizing.

While the experiment was designed to be as fair as possible there were a few slight flaws within it.

The two-stage model used significantly fewer commands on average to traverse the path. The one-stage model totalled 71 commands while the two-stage model totalled 97. A possible explanation for this disparity is that since the one-stage model was hijacked with additional commands more often, these hijacked commands actually took the vehicle closer to the end point. For example, if the user said "go" and then the unintended user said "go" after hijacking a command, then a successful hijack would mean the vehicle would incorrectly travel the length of two "go" commands rather than one. Naturally, because a lot of the commands are repeated, the hijacks actually help the vehicle along quicker. In future experiments, it may be wise to have the hijacker choose a random command to say rather than saying the exact same command as the user. In this way, the hijacked command will only sometimes help the vehicle along its path.

In addition, the tiled flooring wasn't exactly flat (small valleys between each tile). This meant that the vehicle would have inconsistent turns if the wheels got trapped in these "valleys" meaning several "left" or "right" commands would need to be repeated to make a simple turn. This was amplified when the mecanum (360) wheels lost their grip because parts of the floor weren't level.

Another slight oversight was the fact that the left and right paths around the red line on the track do not have equal spacing. Going through the left side has much more space to actually get around the red line. An easier path doesn't necessarily mean a change in the number of commands, but it does mean a chance for flexibility in the type and order of commands available. For example, going around the left will facilitate more "left" commands because it's wider than going on the right.

Doing more trials with the voice models would better provide accurate data and allow confidence intervals to be determined with it. This simply requires more time and additional testing.

This study aimed to apply voice machine learning to real life rather than just comparing loaded wav samples into an online app to test the model accuracy. In real life and outside, there is wind, distance, location, and many other factors which just can't be simulated without the use of a physical vehicle. However, an indoor area was still used for the experiment to maintain experimental rigor while still accounting for a few of the factors of real-life application. The closer testing can get to real life, the better assistive technology can get for the people who need it.

REFERENCES

Bourbakis, N. G. (2008). Sensing surrounding 3-D space for navigation of the blind. IEEE Engineering in Medicine and Biology Magazine, 27(1), 49–55. https://doi.org/10.1109/MEMB.2007.911164

Cook, A. M., & Polgar, J. M. (2015). Assistive technologies: Principles and practice (4th ed.). Elsevier.

Dakopoulos, D., & Bourbakis, N. G. (2010). Wearable obstacle avoidance electronic travel aids for blind: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(1), 25–35. https://doi.org/10.1109/TSMCC.2009.2021255

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. https://www.deeplearningbook.org

Lindström, A. (2024, January 2). *Assistive technology*. World Health Organization (WHO). Retrieved January 13, 2026, from https://www.who.int/news-room/fact-sheets/detail/assistive-technology?

Tapus, A., Mataric, M. J., & Scassellati, B. (2007). The grand challenges in socially assistive robotics. IEEE Robotics & Automation Magazine, 14(1), 35–42. https://doi.org/10.1109/MRA.2007.339605

Warden, P. (2017). *Google Speech Commands*. Kaggle. Retrieved January 13, 2026, from https://www.kaggle.com/datasets/neehakurelli/google-speech-commands

World Health Organization. (2022). Global report on assistive technology. World Health Organization. https://www.who.int/publications/i/item/9789240049451