COMPARING VOICE USER AND GESTURE INTERFACES

A Research Paper submitted to the Department of Computer Science
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By

Vinay Garimella

December 9, 2020

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISOR
Paul McBurney, Department of Computer Science

ABSTRACT

The effect of machine learning on gesture and voice interfacing is an important area of research because of the emergent array of alternative technologies that offer different choices than the more ubiquitous interfaces, like touch screen or mouse and keyboard. Thus, it is essential for researchers to develop a more thorough understanding of the effect of machine learning on gesture and voice interfacing

This study will seek to review the current literature on the effect of machine learning on human computer interaction. Specifically, I will be examining the models used in each of the research experiments and will be comparing the results of the trained models for voice and gesture interfaces. The findings of this study indicate that voice user interfaces are better than gesture interfaces at handling noisy info; voice user interfaces additionally have a higher degree of accuracy than do gesture interfaces. Finally, this review also found that voice user interfaces are more accessible in terms of ease of use, and as such, it is a viable alternative to touch and mouse/keyboard interfaces. Gesture interfaces, however, can also provide a suitable and user-friendly alternative to touch interfaces, provided a straightforward lexicon is used.

1 Introduction

This paper will compare the current state of the art gesture user interfaces and voice user interfaces. The paper will identify the interface and its models that are being used and compare it to the other interface. Currently, human computer interaction is largely touch-based. Therefore, I would like to see if alternate ways of interacting with machines are possible with a degree of accuracy that would make the interface viable to use.

The paper also will compare the results of the models which the corresponding user interface uses. It will not directly compare the models because the different research uses their own models which gives too many variables to compare. By comparing the results of the different models, we can compare how the current state-of-the-art research compares for both voice and gesture interfaces.

Current technologies that use voice user interfaces are Amazon's Alexa, Google's Echo, and Apple's Siri. Similarly, technologies that currently use Gesture Interfaces are Apple's Face ID sensor, as well as Virtual Reality. Because these technologies are starting to be used, I would like to research Gesture and Voice User Interfaces to see how viable it is as an alternative, and possibly a more natural way of interacting with a machine. I will be researching the algorithms and models used to interact with these technologies, and comparing the results.

2 Background

2.1 Gesture Interface

One of the ways that users can interact with machines is using a gesture interface. One of the ways that the user interacts with gesture control. With gesture control, users can communicate with gestures that "carry certain symbolic meanings, a behavior we want to abstract" [7] for interacting with the machine. To recognize gestures, computers need to recognize "dynamic" gestures so that they get the physical context as well as the temporal context, or the context of the gesture over multiple frames. In addition to this, gestures need to be able to handle an array of factors accurately, such as comfort, intuitiveness and lexicon to be useful to users. [7]

2.1.1 Gesture Recognition using mm-Wave sensor for Human-Car Interface

This article researches the implementation of gesture control systems on car "infotainment" [5], which is the combination of information and entertainment systems in a person's car. Due to the increase in complexity of the car's infotainment systems, of which some mainstream operator's offer up to 700 different options, there is a potential increase in driver distraction [5]. This has possibility of increasing the number of accidents and/or deaths on the road.

Due to this Smith, et al. have focused on creating a gesture detection system to decrease the amount of time that the driver has to take their eyes off of the road. This will be based off of a radar system because it is less likely to be affected by other variables such as lighting, color, background and high computational cost [5].

The system design "uses a wireless radar sensor for detection and recognition of gestures" [5]. This offers advantages over other devices such as touchscreen, phone

connection, and buttons and knobs that are currently used in cars. It uses an mm-Wave radar design, which is a "60 GHz frequency modulated continuous wave (FMCW) radar sensor" [5].

Smith et al., used a random forest classifier algorithm for training the data, and used the same for testing the data. The parameters for the data were 10 for the forest size, 10 for the max depth, 10 for the minimum number of samples, 30 for the classifier minimum, and 50 for the classifier buffer size. These parameters were all chosen to get enough tree size and depth for classification while decreasing computational cost [5].

There were multiple sets of sensors that were placed in the car. For each set, the group tested the sensors multiple times. The sensor setup ended up being in the center console as well as the back of the two front seats to make sure the driver and his passengers could use the sensors [5].

The gestures for the first test set were the Low Wiggle, the Turn Over, and the High Grab, which had an average accuracy of 91, 99, and 95% [5]. The gestures for the second test set were the Swipe, the Large Circle, and the Small Circle, which had an average of 91, 97, and 96% [5].

2.1.2 Vision-Based Hand-Gesture Applications

Current user interfaces are normally with touchscreens or mouse and keyboard, which require a learning curve. However, ideally in the future, we will have computer interfaces that do not require such a learning curve.

Wachs goes into different costs and benefits of creating a gesture user interface, for example price has the upside of better equipment with higher price, however if there is not enough budget then the engineer is limited in his applications. The system should be responsive in real time. The system should be able to adapt to user preferences and feedback, and should be easy to learn how to use. The performance of the user interface should have high accuracy, and be able to track the hands and their trajectory.

Wachs says that the gesture interface should be able to take a low mental load to use, which means that users should be able to use the system without much though behind actions. The system should be intuitive to use and should be comfortable for the users and not take much muscle activity. It should be able to take a large lexicon size for both single hand and two hand gestures in addition to recognizing gestures while the user is wearing different types of clothing. Users should be able to reconfigure and personalize their hand gestures and the system should recognize their interaction space. It should also be able to distinguish between different gestures including when the user is not paying attention to articulating themselves.

Wachs states that gesture sensing can be done by gloves, sensors, or cameras, and the data can be trained with programming applications such as machine learning and algorithms. These algorithms have to take in classifications of motion, depth, color, shape, appearance, and multi-cue (which means multiple types of input).

This technology can be used in multiple technologies, such as medical systems and assistive technology, where gestures can be used to control a variety of applications. To entertainment and "crisis management and disaster relief".

For gesture interfacing to be a valid method of user interfaces, Wachs states that there needs to be rigorous validation testing of data through different criteria such as "sensitivity/recall, precision/, positive predictive value, specificity, negative predictive value, f-measure, likelihood ratio, and accuracy." He also adds user independence as well as a qualitative and quantitative assessment as two more categories to test user performance to see if a certain UI model is ready for use.

2.1.3 Machine Learning Based Interaction Technique Selection for 3D User Interfaces

The objective of Lachoche and Duvale's experiment is to "train a machine learning algorithm that can predict the most adapted interaction technique to a particular situation according to the user information" [6]. To increase the ease of use of 3D user interfaces, they can be trained on the preference of each "user's needs, skills, and preferences. This method will be using the 2D Fitts Law approach for the pre-test, and the Support Vector Machines for training the data [6].

Lacoche and Duvale research related work to figure out what they need to build a model. From the results of this work, they find the user profile, the user preferences and interests, the user environment, and the user monitoring information are important properties that must be included in training the model. They also looked at other uses for personalized machine learning, however those were only

used on 2D interfaces, as well as the used as part of window, icon, mouse, and pointer interaction [6].

In order to get the best subjective results, Lachoce and Duvale decided to get the most preferred strategy, and the best performing strategy. The three selection techniques which were compared were the 3D-Ray Casting technique, the 3D Bendcast, and the Occulsion technique. The 3D-Ray Casting technique is when a straight ray is shown on the interface pointing from the user's hand to the object they are pointing to. The 3D Bendcast is when the user interface "automatically bends the closest target" to help the user select the object [6]. The occlusion selection technique is when the user places the tip of the finger between the eye and the object the user wants to select.

Lachohe and Duvale collected data based on physical characteristics, such as gender, handedness, age, master Eye, visual problems, experience with virtual reality, experience with video games, and experience with computer/tablets. They then had the users go through a 2D selection pre-test, and then gave a subjective questionnaire to each user with questions answered on a Likert scale of 1-7 to record the techniques that they preferred.

After recording both the preference and best performance for each of the user setups, the 19.6% of the participants had performances differing based on hardware configuration (Oculus Rift and Zspace), and 64.7% had performances differing based on hardware configuration. For the "consistency between preferred and performances," 49% did not have the best performed technique correlate with the preferred interaction technique in the Oculus interface, and 41% did not have the best performed technique correlate with the preferred interaction technique with the Zspace interface [6].

After recording the data from the pre-test and then recording the preferred and best performing technique on the Oculus and Zspace, Lachoce and Duvale trained the data using a Support Vector Machine classifier. After training the data, they compared the best performing technique and the most preferred technique with the untrained data. For the oculus, the preferred technique was chosen 49.02% of the time, with the profile features the preferred method was chosen 57.17% of the time after training on the SVM, and with the pre-test features the model chose the preferred method 55.6% of the time trained on the SVM. For the Zspace, the best performing technique was chosen 43.14% of the time, with the profile features the best performing method was chosen 52.95% of the time after training on the SVM, and with the pre-test features the model chose the best performing method 52.95% of the time trained on the SVM [6].

2.1.4 Hand Gesture Recognition in Automotive Human–Machine Interaction Using Depth Cameras

Zengeler et al. are experimenting with using cameras to create a gesture user interface for cars. Because humans already communicate non-verbally through context-sensitive hand gestures. To achieve this, they decided to use time of flight sensors which can convert depth measurements into depth-data input measurements in real time. They will then load this data into a mobile tablet which will run an infotainment application on a mobile tablet [7].

They first gathered data from a hand posture dataset from a previous experiment which consisted of 600,000 images of 20 different people, and 450,000 images of 15 different people. They then used Principal Component Analysis (PCA) to break down their data into the necessary components. They then used multiple different methods to train their data, a multi-layer perceptron (MLP), a SVM, a convolutional neural network (CNN), a two-layer MLP (with the output of the first one feeding into the second), and a Long Short-term Memory (LSTM) Network.
The MLP had a training performance of 93.7% and a test performance of 98.7%, The MLP had a training performance of 93.7% and a test performance of 98.7%. The SVM had a test performance of 99.8%, The two stage MLP had a training performance of 97% and a test performance of 97%, The CNN had a training performance of 94.5% and a test performance of 94.5%, and the LSTM had a training and test performance of 100% [7].

2.2 Voice User Interface

Another way that users can interact with the machine is the voice user interface. They allow users to interact by measuring the type and intent of words and turning those into actions in the computer. Different types of models and algorithms that are used in Voice User Interfaces are natural language processing, WaveNet, and hierarchal models [1,2,4]. This allows for a user to have a eyes and hand free way to interact with their machine.

2.2.1 Patterns for How Users Overcome Obstacles in Voice User Interfaces

Due to Voice User Interfaces (VUI) growing in popularity, Myers et al. want to implement a VUI calendar system to analyze how different tactics are used to navigate obstacles with VUIs. They tested users, all of who have testified as having technical backgrounds to answer to questions, what categories of obstacles did users encounter, and what tactics did they develop to counter these obstacles [1].

To test this the used a NLP library which would process voice commands. They tested a few obstacle categories, unfamiliar intent, NLP error, failed feedback, and system error. These errors occur if the user gives a command which the NLP cannot parse, if the NLP mishears a command, when users ignore or mishear feedback, or a flaw in the VUI system architecture respectively. After the users tested the VUI, Zengeler et al. found 10 common tactics that most of the users tried, such as hyper articulation of commands, simplification of commands, giving a new command, add additional info, rely on the graphical interface, accepting a wrong response from the VUI, restarting from before they encountered the obstacle, repeatedly repeating commands in frustration, quitting the task in frustrating, or recalling the correct utterance without help. Of the tactics, hyper articulation was used the most at 40.48% of the time, and Hyper articulation, additional info, simplification, and new utterance were the top 4 at 77.46% of tactics tried [1]. The rest of the tactics were fallback tactics after trying the initial tactics.

Zengler's experiment found that 52.1% of the common errors that were found were in NLP errors. However, they decided that this was not the biggest threat to implementing VUIs in UX designs, rather the other obstacles were the ones that caused the most frustration for the users. They wish to continue this study with a larger sample size and a more diverse demographic group as well as with "non-technology comfortable users" [1].

2.2.2 WaveEar: Exploring a mmWave-based Noise resistant Speech Sensing for Voice-User Interface

Due to the noise that a VUI encounters from outside interferences, Xu et al. are exploring a method of creating a model, which they call WaveEar, which is an "end-to-end noise resistance speech sensing system" [2]. The outside interferences come from audible interference from daily environments, such as buses and other transportation, in addition to sever Denial-of-Service (DoS) style attacks.

The mm-Wave style of sensing voice commands is that it is Noise-resistant, informative, and secure. However, unlike microphones mm-Waves are unidirectional, i.e they only sense sound in one direction. From there they develop the WaveEar. It does this by training the data in an encoder and decoder neural network which encodes 3 convolutional layers and decodes 2 convolutional layers. It then constrains and minimizes the loss through mean squared error (MSE) and back propagation.

The experiment was run with 21 users, 11 male and 10 female which included both native and non-native English speakers. One evaluating metric were the speaking signal to noise ratio, which calculated the amount of noise that detects the vocal pressure, but not the sound pressure. The second evaluating metric was the sensitivity. The third was Word Error Rate, which measures the difference between the reference and recovered words. The last evaluation metric was Mel-Cepstral distortion which evaluates the voice based on the human acoustic system preference.

The results indicate that the device only introduces mild sound distortion. The device takes the analyzed signal and recognizes it through with a <0.01% training and test loss, and an average accuracy of 96.7% [2]. It can also map the signals to speech directly. After testing the device with ambient sound such as pop music, presentation, and water flow, they propose that the WaveEar is immune from ambient noise. The WaveEar is also "robust to users' different emotional state" which means that it can handle changes in pitch and volume due to emotion. The mmWave does have a small issue when there are a lot of people in the same area, though this can be mitigated by directional targeting of the user.

2.2.3 Conditional Driving From Natural Language Instructions

The goal of Junha Roh is to implement a VUI for car-human interfacing due to the upcoming prevelance of self-driving cars. They implement a hierarchical model with implements a encoder called a Gated Attention model which creates 50 embedding vectors fed into a GRU to encode the single instruction. This gated attention model is fed into the lower level model which uses a series of convolutional and ELU layers to decode the image. The car

then takes the image and command and uses it to determine the correct direction to go.

The training data is obtained by sending out a romaing expert into the simulation and then obtained the sub-task labels based on the road map [4]. Then they "partitioned the trajectories into a set of trajectory snippets" such as left, right, straight, and lanefollow.

The results of the training data were a max of 100% and 100% for train and test for single commands, 100% and 98.6% train and test of double commands, and 100% and 93.8% train and test for ordinal commands. For all of them combined the best train and test results were 100% and 97%. Between the model and the two-hierarchal variants, the biggest difference between the models is the "end-of-sub-task" value that was generated from the model rather than the models themselves.

For future work they do note that work on ambiguous speech needs to be done, because most speech are not clear commands.

3 Comparing Gesture Interfaces and Voice User Interfaces

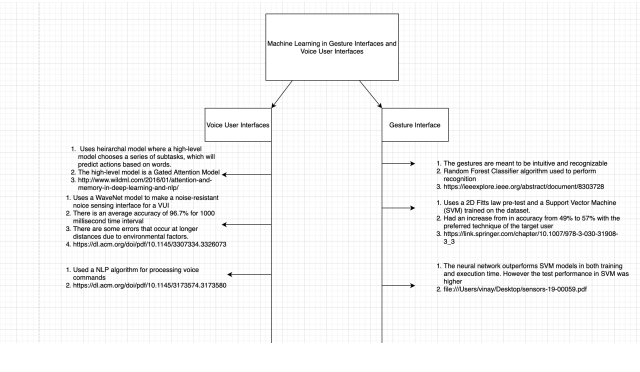

Figure 1: This figure highlights the different machine learning models used in Voice User Interfaces and Gesture Interfaces

There is some variable accuracy that the gesture interfaces come across with, with higher degree of range and movement increasing the amount of accuracy of which a machine can interpret a gesture [5]. To get a more robust system, multiple users must use similar gestures to give the random forest algorithm more variations which will allow the algorithm to ensure "usability and robustness" [5]. This is because with the random forest algorithm increasing the number of decision trees will increase the accuracy of the system.

Similarly, voice-user interfaces can handle noise as well, but their design may be a bit more obtrusive [2]. For the machine to decrease noise, it needs to handle the noise in the model, as well as cancel noise interference externally as well. This is possible through "active noise cancellation" [2] where devices can send out sound waves which cancel out the incoming noise. The mmWave sensor [2] is able to sense the noise resistant speech of the user and accurately represent the sound waves as speech. It does this by training the data in an encoder and decoder neural network which encodes 3 convolutional layers and decodes 2 convolutional layers. It then constrains and minimizes the loss through mean squared error (MSE) and back propagation. The device takes the analyzed signal and recognizes it through with a <0.01% training and test loss, and an average accuracy of 96.7% [2].

However, this means that voice user interfaces need to have obtrusive machinery to drown out extra noise [2]. The gesture recognition has the possibility of using a camera system to get around an invasive interface [7]. However, some gesture interfaces require the use of headsets and large pieces to wear on the user's body [6]. The invasiveness of this type of machinery can make it exhausting the wear and manipulate the interface, as well as restrict the applications of it. The invasiveness of the noise cancelling machinery can cause irritation by prolonged skin contact [2]. The WaveEar model does mitigate the need for this using a localized sensor to gather the data, and a noise reduction model to reduce noise. Due to the amount of machine on the person being smaller this allows for the person to use a VUI without the invasiveness of the machinery. This, in certain situations, offers the VUI an advantage over the GUI by allowing it to be a less invasive experience. By being a less invasive experience, it will appear to the users more by being a more natural way of interacting with their devices and other people through those devices.

When it comes to the sensors, certain sensors can decrease the accuracy of the results. "Camera based sensors are susceptible to changes in light, color, background, and have high computational costs" [5]. They also have to be able to handle large lexicons of physical gestures, which usually underperforms datasets trained on a smaller sample size of lexicons [3]. While depth-based sensors can have trouble recognizing a hand gesture due to the amount of dynamic hand movement [5]. Similarly, in a voice user interface, articulation, volume and accent can all give enough variability to confuse the machine enough to misinterpret a command [1]. This could lead to the users getting frustrated with the machine and possibly giving up on correcting it [1].

This means that both the VUI and GUI have the possibility of frustrating their users when the machine misinterprets their commands. This can be corrected manipulating or changing the models until the accuracy changes. Some ways

to do this are to increase the number of epochs the models train on and changing other hyperparameters to see if they can better tune the model.

Another problem that a Voice User Interface has is outside interference from multiple people being in the same area. The WaveEar sensor can isolate a "speaker among multiple people" [2]. One way that the Voice User Interface can use to reduce outside interference is to use an mmWave sensor. The mmWave sensor allows the WaveEar model to ignore real-world noise [2]. This allows for a person to communicate with their device inn an area with a lot of noise. This is an advantage for using voice commands because it allows for the user to have a non-invasive and low noise method. In the current research into the state-of-the-art for Gesture Interfaces, it has been found that "depth from stereo is usually coarse-grained and rather noisy" [3] because of this it is combined with other cues to reduce the noise. In addition, Lachoche [6] touches upon this point for how it would be interesting to study such a scenario in a noisy environment.

One obstacle that Gesture Interfaces have over Voice User Interfaces are that they have to recognize take depth and a wide array of angles into account. This was shown when "fusing stereo camera data from depth sensor" to a Time of Flight sensor [7] which increased the accuracy of the gesture recognition. For example, in a car the sensor where the "majority of the radar beam spread into free space" to detect objects and unimportant objects. After that they added sensors for the driver and passengers [5]. This means that to accommodate the users they need to be designed around environmental "constraints" [5]. However, with voice user interfaces, the interference comes from outside noise is more of a prevalent issue [2]. This means that while noise is a bigger issue for the voice user interface the range of lexicons are a bigger issue for the graphical user interfaces. Graphical user interfaces also have to deal with ease of use and learnability [3] when it comes to this lexicon due to more people having a knowledge of spoken language over sign language. This means however the machine is able to recognize hand gestures from the user it needs to be in a way with is easy for the user to do for a long period of time and as something the user does not have to relearn. This is a major advantage for the advocating for voice user interfaces because it means that the users do not need to learn a new gesture or language to interact with the machine. The lower learning curve can make it more likely for people to use it.

Voice User Interfaces in cars can allow for the driver to control their car through speech [4]. It uses a hierarchical model with the higher level being a Gated Attention model, and then applies trains it on convolutional, ease of language understanding (ELU), and gated linear units (GLU) [4] in the lower model. The GUI for the car uses a CNN with two convolutional layers from with the output is the input for an MLP [7]. The interface has an accuracy of about 98.2 percent for single commands, while having lower score of 97.6 percent for double commands. This was using a cross-entropy loss for training the model [4]. As stated before, the gesture user interface uses a random forest model [5], the random forest model gives an accuracy from 91-99% for different gesture types, with the smaller range gestures having a higher degree of error [7]. This means that both the Voice-User Interface and the Gesture Interface have high degrees of accuracy, but that the voice-user interface is safer to use. This is because the voice user interface has a similar and higher accuracy than most of the accuracy rates for the gesture user interface. The range of accuracy for the GUI is concerning because in certain situations the computer is more likely to not recognize the user's intentions which can cause an accident. There is also the possibility of the driver making unclear gestures due to the more forgiving nature of making bigger gestures in the car, as it correlates with less error. To move forward with using gestures in the car, there needs to be much less error [7].

This could be mitigated by having all of the gestures be larger, which will raise the accuracy of the results due to the model being better at recognizing such actions. However, this raises the risk of fatigue for drivers, as using large gestures for a long period of time or during heavy traffic can get tiring. Driver fatigue can cause them to make more errors in their gestures raising the risk of accidents. For now, gestures and voice commands can be used for less important tasks, such as changing the radio, because it has a much smaller risk being more likely to annoy the user if it does not work rather than cause a car crash. In contrast voice commands cause less fatigue, but can cause error if the user does not speak loud enough.

6 Future Work

In the future I would like to recreate some of the experiments used in the cars, and do a direct comparison of both the gesture and voice user interfaces within the same experiment. I believe that gestures and voice commands can be a great improvement for the driving experience if done correctly. For each of the interfaces I would train multiple models such as CNN, MLP, SVMs etc. and compare the best model for each interface. I would like to get an error rate of less than 0.1%, which I believe is a sufficient rate for these interfaces to be implemented in cars with the current method of driving as a backup. This will allow users who need a break the peace of mind that they will not crash while taking their hands of the wheel. I believe that this will give a much better comparison between the two different interfaces.

The criteria that I would use to compare the two would be the accuracy and precision from the models. By having both interfaces be tested in the same experiment, I could have a direct comparison in testing which type of interface gives better ease of use and a more natural experience. By comparing these two factors it would be possible to find which interface is safer to use and which interface is easier to use. Afterwards I would like to see the results of an interface which use both gesture and voice commands to interact with the computer. By looking at the accuracy and ease of use from this experiment, we can see how much better or worse it is to use both interfaces.

In addition to this I would like to research different types of models from Lachoche's experiment [6] to see if I could improve on top of his results in discovering the best performing technique. I would try to do this experiment with their SVM model and see if increasing the size and adding any layers to the model would increase the accuracy of the results.

5 Conclusion

Comparing the voice user interfaces and gesture interfaces, I found that both types of interfaces had their own obstacles to overcome. Because there is not a completely homogeneous environment in which a person can use a machine in the real world, there will always be some type of noise that these types of interfaces have to overcome. The different interfaces handle noise in their own ways, relying on their own methods to reduce noise, and they both do so to a high degree of accuracy However, voice user interfaces tend to have a higher degree of accuracy than gesture interfaces, as well as a lower loss rate.

The Gesture Interface has to accurately detect a movement from a certain range and movement. This requires the use of multiple sensors which will have to be able to calculate the depth, shape, and movement of an object with outside error. The gesture interfaces achieve recording a dataset of lexicons with different hand and body gestures. To accurately record these gestures, the device has multiple cameras which will capture the depth, color, shape, and other features. The models, such as random forest, will then train the data.

The Voice User Interface has to accurately detect speech patterns and reduce the noise from outside interference. The voice user interface has the ability to counter noise through the model and through the machine as well. It does this by including noise cancelling as part of the device and using math to reduce loss while simultaneously training the model. It achieves this level of accuracy and loss rate by using models such as Wave-Ear [2]. Another use of the Wave-Ear allows the VUI to cut down on outside interference without bulky and invasive machinery.

The Voice and Gesture interfaces both reduce noise well; indeed, voice and gesture interfaces could become staples in the future, given the direction in which the technology is advancing. In the meantime, touch and mouse and keyboard interaction in the computers are the norms.

Additionally, another promising use for the voice user and gesture interfaces is in cars. In motor vehicles, the interfaces had high degrees of accuracy with recognizing speech and gestures respectively, however the gesture recognition ended up having a higher range of error, which is dangerous to the people who are on the road due to the potential of lost lives. I believe that both interfaces are not accurate enough for them to be used as a replacement for driving, but the gesture controls are far behind the voice controls due to the amount of error that some gestures have.

REFERENCES

[1] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*)*. Drexel University, Philadelphia, PA, 7 pages. DOI: https://doi.org/10.1145/3173574.3173580

[2] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, & Kun Wang, Wenyao Xu. 2019. WaveEar: Exploring a mmWave-based Noise-resistant Speech Sensing for Voice-User Interface. In the 17th Annual Int'l Conference on Mobile Systems, Applications, and Services (MobiSys '19), June 17–21, 2019, Seoul, Republic of Korea. ACM, NY, NY, USA, 13 pages. https://doi.org/10.1145/3307334.3326073

[3] Juan Wachs, Mathias, Kölch, Helman Stern, Yael Eden. 2011 Vision-based hand-gesture applications. *In Communications of the ACM, 12 pages* https://dl.acm.org/doi/pdf/10.1145/1897816.1897838

[4] Junha Roh, Chris Paxton, Andrzej Pronobis, Ali Farhadi, Dieter Fox. 2019. Conditional Driving from Natural Language Instructions. In *3rd Conference on Robot Learning (CoRL 2019), Osaka, Japan, 12 pages* http://proceedings.mlr.press/v100/roh20a/roh20a.pdf

[5] Karly A. Smith, Clement Csech, D. Murdoch and G. Shaker, "Gesture Recognition. Using mm-Wave Sensor for Human-Car Interface," in IEEE Sensors Letters, vol. 2, no. 2, pp. 1-4, June 2018, Art no. 3500904, doi: 10.1109/LSENS.2018.2810093.

[6] Lacoche J., Duval T., Arnaldi B., Maisel E., Royan J. (2019) Machine Learning Based Interaction Technique Selection for 3D User Interfaces. In: Bourdot P., Interrante V., Nedel L., Magnenat-Thalmann N., Zachmann G. (eds) Virtual Reality and Augmented Reality. EuroVR 2019. Lecture Notes in Computer Science, vol 11883. Springer, Cham. https://doi.org/10.1007/978-3-030-31908-3_3

[7] Nico Zengeler, Thomas Kopinski, Uwe Handmann. 2018. Hand Gesture Recognition in Automotive Human-Machine Interaction Using Depth Cameras *Understanding Policy-Based Networking* (2nd. ed.). University of Applied Sciences, Bottrop, Germany