Android Speech Interface to a Home Robot

July 2012

Deya Banisakher
Undergraduate, Computer Engineering
dmbxt4@mail.missouri.edu

Megan Biondo
Undergraduate, Computer Science
mmbvfb@mail.missouri.edu

Tatiana Alexenko
Graduate Mentor
ta7cf@mail.missouri.edu

Prof. Marjorie Skubic
Faculty Mentor
SkubicM@missouri.edu

Abstract

A growing elderly population and shortages of nursing staff create a need for innovative technologies in eldercare. The use of a home robot to assist with daily tasks, such as fetching objects, is one such example. However, there is also need for effective and convenient human robot interaction in this scenario, which a simple speech interface may provide. We investigated the use of the built-in speech recognition in Android phones for the fetch task, as well as the various methods of implementing a successful and efficient two-way server-client connection over an appropriate and practical type of wireless network between a smartphone and a home robot. An Android application was developed which utilizes the underlying network and process communication system to support its use. Finally, tests were performed comparing the accuracy of speech recognition on the Android phone for older and younger adult voices.

Introduction

Recent studies have shown that one of the top five tasks noted by seniors for assistive robots is help with fetching objects, for example, retrieving missing eyeglasses (Beer et al., 2012), and the preferred form of communication with the robot is a speech interface (Scopelliti et al., 2005). We investigated the use of the built-in speech recognition in Android phones for use in this scenario. We created an Android

application and implemented the underlying network and process communication system to support its use. We also collected voice recognition transcriptions from old and young people; they spoke into an Android device that had a testing application installed which we have developed. We also compared the accuracy of speech recognition on the Android phone for older and younger adults, as well as male and female ones.

Previous Works

Skubic et al. have studied spatial language in older and younger populations. In collaboration with Carlson et al. at Notre Dame Dept. of Psychology, they collected speech samples of older and younger adults giving spatial descriptions (Carlson et al, in review). They also created a robot capable of recognizing furniture and processing textual spatial descriptions, in addition to the common robot capabilities such as obstacle avoidance. The robot was made to listen to commands coming from the user through a computer's keyboard that is wired to the robot itself. Since it is impractical to type the spatial descriptions, there is a need for an accurate speech recognition which we addressed in our research.

Why Android?

We decided to test Android's speech interface, created by Google, because it is known for high accuracy and is freely available in Android-based devices which are being activated at a rate of 1 million devices per day worldwide (Android, 2012). Google's approach to speech recognition is also unique because it relies on crowd-sourcing in addition to integration of existing acoustic models.

We created an Android application that handles the audio data and sends the transcription to the robot for processing.

The use of Android devices for this purpose also has technical benefits including the audio processing and transcription is handled by Google's servers, Android applications are easy to install on any Android device, Android devices and the operating systems support a wide range of accessibility features for helping the elderly use the different applications installed, Android devices have built-in microphones, eliminating the need for the user to purchase a headset or other microphone, and finally a speech recognition application allows the user to decide when they want to communicate with the robot, which prevents the robot from reacting to speech directed to other people.

System Components

The system as a whole consists of two main components, an Android phone and a robot. Both components interact and send information to one another using a specific networking algorithm, Transmission Control Protocol (TCP).

Android is based on Linux kernel which is an open source base for the growing operating system, it also utilizes Java's API into its development which allows it to perform and function in an object oriented way. Furthermore, Android, when it comes to development, makes it easy and practical for developers to change, switch and supply more resources to their applications by dealing with the XML based resources. XML is a simple language that Android allows developers to use to create and reference to sophisticated screen layouts and other resources such as pictures and videos.

Android's platform and its use of Java's packages such as java.net, allows developers to use the phones hardware in a matter that is no different than a one in a fully featured computer. The networking

capabilities of the Android phone leave the users with the freedom to choose which networking protocol they would like to follow and integrate in their applications.

"ROS [Robot Operating System] is an open-source, meta-operating system for your robot. It provides the services you would expect from an operating system, including hardware abstraction, low-level device control, implementation of commonly-used functionality, message-passing between processes, and package management. It also provides tools and libraries for obtaining, building, writing, and running code across multiple computers" (ROS, 2012).

Ultimately, for the purposes of this undertaking, ROS is a tool that can be used to program and control a robot. The robot uses ROS which is based around publish-subscribe pattern. The server process inside of ROS publishes the textual transcriptions it receives from the Android device while other processes in the robot (primarily language processing) subscribe to the server's feed. In order for the robot to receive these transcriptions, a TCP server had to be integrated into ROS.

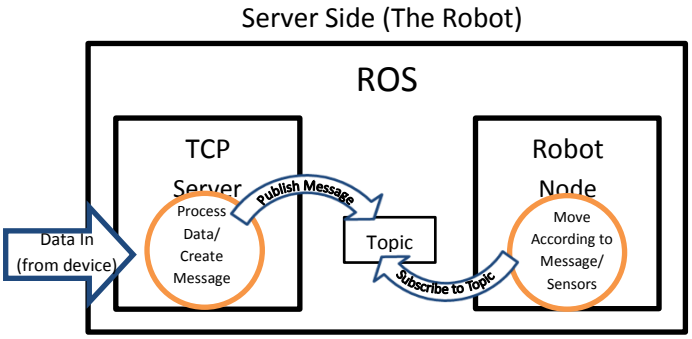
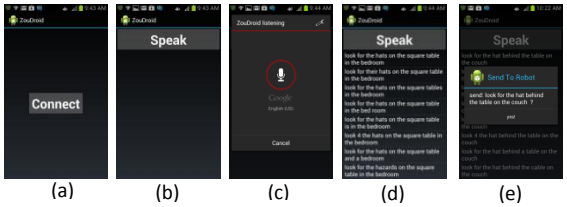


Figure 1. Server communication within ROS.

There are many ways to achieve a link between two devices, but for matters of reliability a TCP server used. TCP, or Transmission Control Protocol, uses what is called a sliding window to assure that all packets reach their destination. Although large amounts of vital data are not being sent, in this case of something as simple as a sentence or two, it is important that all the pieces make it to the destination.

System Functionality

Everything begins when the user decides they want to use the robot. When they open the application and begin, networking is established. The user speaks into the android device, and then the phone will connect to Google's voice engines and obtain a set of transcriptions. The user, if satisfied, with any of the transcriptions, selects to send the transcription to the robot. The phone will prompt the user to confirm their option and send to the robot. The use, if sure, will accept and the transcription will be sent to the robot as shown in Figure 1.



(a) (b) (c) (d) (e)

Figure 2. (a) User connects to robot. (b) User chooses to speak into phone. (c) User speaks into phone. (d) Phone displays the possible transcriptions to user. (e) Phone prompts user to send selected transcription to the robot.

Upon receipt of the transcription from the Android phone, the server, which is part of the robot, stores the transcription temporarily, and then it sends the Android Phone a message stating that it "got it." It will then send the transcription to the other nodes on the robot to be processed.

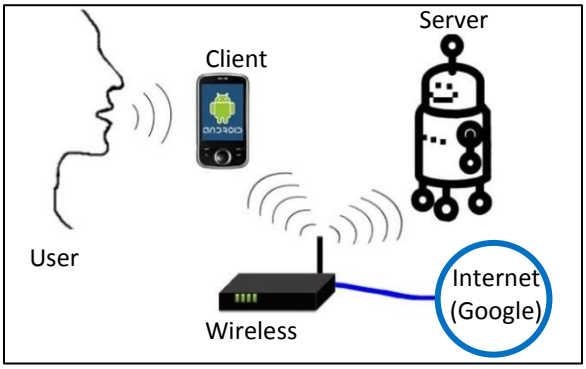


Figure 3. View of overall system communication.

Testing Methods

Originally we used prerecorded statement from both young adults and older adults. Recordings were played through a speaker and directed at the android phone. After a total of 29 transcriptions were taken from 16 different older adult voices, it seemed apparent that this method was not effective as the phone did not recognize any one sentence in its entirety. In fact, it recognized less than half

of the sentence for the majority of those recordings. The recordings of the younger group were better but the phone still only recognized very few (13 out of 49). This however, was not necessarily a result of poor speech recognition but a consequence of using recordings, rather than live human voices. One major issue with this method is that there is a loss of quality and clarity in the audio as well as electronic interference, which could alter the outcome of the transcription. So it was decided that we must recruit people to speak into the phone.

We were able to obtain 53 transcriptions from older adults and 48 from younger adults. It was immediately obvious that this was a much better method of testing the voice engine. The subjects were given descriptions to read from the recordings used earlier. These descriptions ranged from a length of 11 to 28 words. After collecting the data we calculated the accuracy of each transcription as well as a binary value. Accuracy was computed by dividing the number of correctly transcribed words by the total number of words spoken while the binary value simple represented rather or not the sentence was transcribed perfectly.

Results

We found that there was a difference of approximately 10% between the average accuracy of transcriptions of older and younger voices with the younger voices being better.

(a)

Younger Adult Voices	# Trans.	Average	Std. Dev.	Min.	% Perfect
Men	28	94.25%	9.69%	66.67%	60.71%
Women	20	90.18%	14.67%	37.50%	40.00%
All	48	92.55%	12.05%	37.50%	52.08%

(b)

Older Adult Voices	# Trans.	Average	Std. Dev.	Min.	% Perfect
Men	22	79.25%	15.86%	42.86%	9.09%
Women	31	84.66%	16.96%	16.67%	32.26%
All	53	82.41%	16.58%	16.67%	22.64%

Figure 4. (a) Accuracy results of younger adult voices. (b) Accuracy results of older adult voices.

Also according to other research there is seems to better recognition of elderly female voices than elderly male voices (S. Anderson et. al., 1999). We had similar results but also noticed that within the younger group, male transcriptions were actually more accurate. These findings not only appeared in the accuracy percentage but also shined through in the number of perfectly recognized sentences of each group. Although it seems that Google's voice recognition overall is reasonably effective.

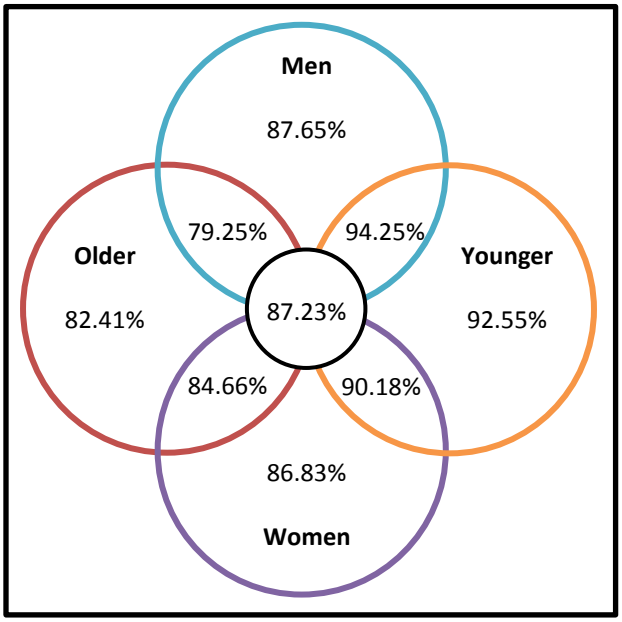


Figure 5. (a) Accuracy of speech recognition comparing older adults vs. younger adults as well as male and female.

However, for the purposes of the fetching goal, there needs to be more structure around the results, such as which words are important and which are not needed at all. For instance, some people may give a very detailed descriptions but the robot is only going to pick out certain words. So if the speech-to-text is mostly have trouble with words such as "it", "the" , and less important words perhaps even some low accuracy transcriptions will still be functional.

Conclusion

In this paper, we have researched Android's Networking capabilities and the accuracy of Google's voice recognition engine. An Android application was developed using Android's API libraries. The application was designed to listen to the user's commands and access Google's engines through a wireless router that is connected to the internet in order to obtain

a set of transcriptions. An algorithm was written and coded to use one of Android's networking features, TCP networking, to send the desired transcription to the server, the robot, wirelessly through the router's local network. We also obtained and compared live transcriptions from both old and young adults in order to investigate the voice recognition engine's accuracy. The research results have shown the effectiveness of the networking algorithm developed alongside Android's networking features. Moreover, the results have shown a significant difference between the transcriptions' accuracy for old and young adults favoring the young male ones. The results can be justified by the fact that Google's unique engines rely on crowd-sourcing, and one can comfortably argue that young adults are the higher suppliers of audio recordings to Google's engines and acoustic models. Overall, though, the results stated in this paper support the effectiveness and accuracy of Google's engines in transcribing voices over the cloud.

References

1. Android, 2012. Android, the world's most popular mobile platform. http://developer.android.com/about/index.html
2. Beer, J.M., Smarr, C., Chen, T.L., Prakash, A., Mitzner, T.L., Kemp, C.C. & Rogers, W.A. 2012. The domesticated robot: design guidelines for assisting older adults to age in place. In Proc., ACM/IEEE Intl. Conf. on Human-Robot Interaction, 335-342, March, 2012, Boston, MA
3. Carlson, L., Skubic, M., Miller, J., Huo, Z., and Alexenko, T. In Review. Investigating Spatial Language Usage in a Robot Fetch Task to Guide Development and Implement of Robot algorithms for Natural Human-Robot Interaction. *Topics in Cognitive Science*.
4. ROS. KenConley. 02 March 2012. 21 June 2012. http://www.ros.org/wiki/ROS/Introduction
5. S. Anderson, N. Liberman, E. Bernstein, S. Foster, E. Cate, B. Levin, and R. Hudson. 1999. Recognition of elderly speech and

voice-driven document retrieval. In *Proceedings of the Acoustics, Speech, and Signal Processing, 1999. on 1999 IEEE International Conference - Volume 01* (ICASSP '99), Vol. 1. IEEE Computer Society, Washington, DC, USA, 145-148. DOI=10.1109/ICASSP.1999.758083 http://dx.doi.org/10.1109/ICASSP.1999.758083

6. Scopelliti, M., Giuliani, M., and Fornara, F. 2005. Robots in a domestic setting: a psychological approach. *Universal Access in the Information Society*, 4(2): 146-155.