

A Comparison of Commercial Speech Recognition Components for Use in Police Cruisers

Brett J. Vinciguerra and Andrew L. Kun

Electrical and Computer Engineering Department, University of New Hampshire

ABSTRACT

The Project54 system seamlessly integrates control of all in-car electronic police equipment into one standard interface. The Project54 system can be completely controlled by voice commands. The two goals of the project are to improve officer's in-vehicle safety and integrate control of all in-vehicle electronics. Since speech recognition in the Project54 system plays such an important role in officer safety, it is crucial that it functions as well as possible. The main goal of this research was to accurately predict which commercial speech recognition components would perform best in a noise robust police cruiser environment. Rather than the traditional method of evaluating speech recognition components solely on accuracy, an emphasis was placed on the effect that falsely recognized words would have on the efficiency of the police officer using the system while driving.

INTRODUCTION

Project54 [1] is a collaborative research and development effort between the University of New Hampshire and the New Hampshire Department of Safety. The Project54 effort concentrates on the integration of embedded mobile computing equipment and wireless networking into NH State Police cruisers. The Project54 system integrates car-based electronic systems to offer advanced support for NH State Troopers.

The Project54 system allows police officers to control devices such as lights, sirens, radar units, police radios, global positioning systems, as well as access record databases, through one standard user interface. The user interface allows input using a touch screen, a keyboard or voice commands. Controlling the system by voice allows for the officer to keep his or her eyes on the road and hands on the wheel while driving. For this reason, most of the interaction with the Project54 system will be done through voice commands from the officers using the system. Since police officers are going to be using voice commands to perform tasks such as turn on lights and sirens before crossing a busy intersection in a high-speed pursuit it is critical that speech recognition (SR) components perform reliably.

The SR portion of the Project54 system is modular and allows any microphone and SAPI 4.0 compliant SR engine combination to handle voice commands. Of the

available off-the-shelf components, four SR engines and four microphones have been selected for possible use in the Project54 system. It is important to determine which combination of these SR components will be more accurate and efficient than the other combinations when used in a police cruiser environment. Accuracy is the number of correctly recognized phrases divided by the number of phrases spoken. Efficiency is calculated using a weighted average that takes into account the amount of false recognitions and how they will impact the end-users of the Project54 system.

In order to determine which components when combined will be more accurate and efficient as part of the Project54 system, the SR components under consideration need to be evaluated in the actual police cruiser environment. However it is not practical to evaluate every SR component combination on the road with officers on duty. This would be very time consuming and it would be impossible to ensure that all the components were tested under the exact same conditions. For this reason the SR components under consideration for use with the Project54 system were put to a head-to-head test in a simulated police cruiser environment using the Project54 Speech Recognition Evaluation Program (SREP).

BACKGROUND

Although accuracy ratings are available for every component under consideration, they do not provide enough information to accurately predict their performance with the Project54 system. The tests that are performed to determine speech recognition engine accuracy ratings are usually performed differently for different engines and they are usually done in an office environment that is much different than the police cruiser environment.

AT&T Bell Laboratories presented the first successful speech recognizer in the early 1950s. It was a hardware based speaker-dependant recognizer that recognized digits from zero to nine. Since this demonstration, researchers have tried to find an accurate method of measuring the performance of speech recognizers. Bell Laboratories claimed an accuracy rating of 98% on their first recognizer. This may seem far-fetched since modern day speech recognizer manufacturers still claim an accuracy rating of around 98%. However 98% accuracy was being publicly demonstrated in the early days of speech recognition [2]. These high rates were

achieved by using trained speakers, or the researchers themselves, who knew exactly what the recognizer wanted to hear. The speakers trained the recognizer with speech patterns that they could easily reproduce during demonstrations. Seeing these results in 1969 prompted S R Hyde of the Joint Speech Research Unit to state 'The accuracy of speech recognizers is 98%' which was followed by 'Because speech recognizers have an accuracy of 98%, tests must be arranged to prove it' [3]. This tongue in cheek statement is referred to as Hyde's Law.

Hyde's Law points out that accuracy ratings of speech recognizers depend greatly on the method in which the recognizers are tested. Since there is no standard method for testing the performance of speech recognizers, published results can not guarantee the actual performance of a speech recognizer in a particular environment except under the exact conditions it was tested under. In order to make a meaningful prediction on which speech recognizer will work the best for a particular application, it is important to evaluate the recognizers in question side by side in the actual environment over a long period of time [4]. Most often this is not possible, so the next best option is to evaluate the recognizers under conditions that simulate specific conditions from the operating environment.

Although microphones are usually tested in a standard way, the results do not provide information about how they will interact with specific speech recognition engines or computers.

HARDWARE AND SOFTWARE

The SREP was implemented on a desktop computer running the Project54 software. A block diagram of this process is shown in Figure 1.

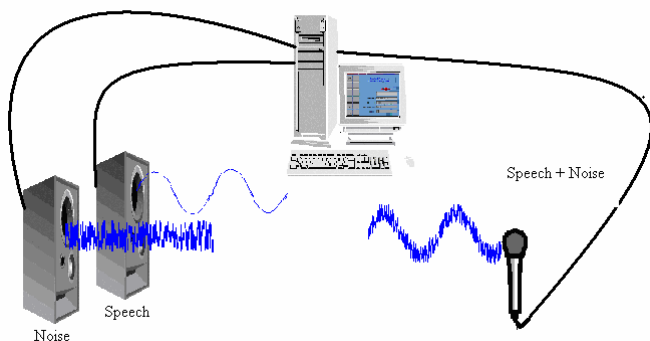


Figure 1 System Level Block Diagram

The computer was equipped with two soundcards. One soundcard was used to output speech commands over one speaker while the other soundcard was used to output background noise over another speaker. Controlling two sound cards made it possible to use one computer that output both the speech commands and background noise at the same time over different speakers as well as "listen" with a microphone to pick up

this combined sound as input to the SREP. The SREP then determines if the phrase recognized through the microphone matches the phrase that was played out of the speech command speaker.

The Project54 system software is completely modular and is based on Microsoft's Component Object Model (COM) [5]. At the center of the Project54 system is the Application Manager. The Application Manager coordinates and sends control messages to the individual applications that control each electronic device in the police cruiser. A visual representation of the Project54 software architecture is shown in Figure 2 [6].

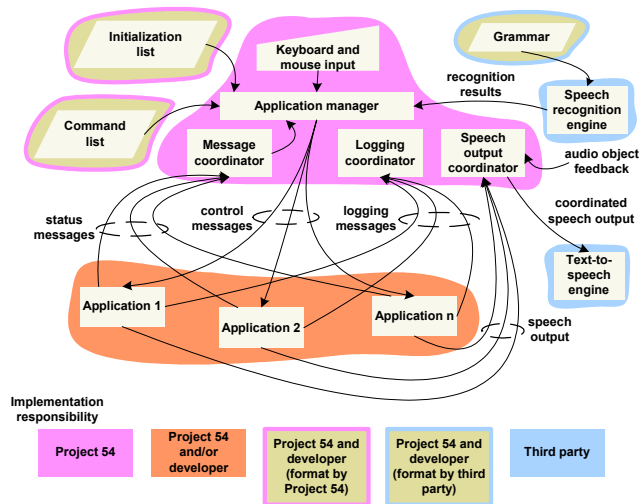


Figure 2 The Project 54 System Software Architecture

All speech input is handled exclusively by the Application Manager, which interacts with the speech engine and sends recognized input to the corresponding application. In order to accurately evaluate the performance of speech recognition in the Project54 system, the SREP was designed in the same way as all the other applications in the system. This insured that speech would be handled in the same way in the tests, as it will be in the actual system.

TESTING

Four microphones and four speech recognition engines were used with the SREP to create 16 unique combinations to be evaluated. The four microphones will be referred to as Microphone A, B, C, and D. Microphones A, B and C are array microphones. Microphone D is a directional microphone [7].

The four speech recognition engines that were evaluated will be referred to as SR Engines 1, 2, 3 and 4. SR Engines 1 and 2 are speaker independent continuous recognizers. Engines 3 and 4 are speaker dependent continuous recognizers.

In order to insure that varying SR engine parameters, the default settings were used for each individual engine. For the two engines that were speaker dependent, the absolute minimum amount of required

training was performed prior to evaluation with the SREP. Each person, whose speech recordings were used by the SREP, created a user file for that specific engine. The training material used was provided with the speech recognition engines and did not include the phrases that were used for testing with the SREP.

The microphone volume was set once for each microphone prior to testing using the Microsoft Microphone Setup Wizard. The Microphone Setup Wizard automatically adjusts the microphone volume to the level it expects to result in the best performance in the current environment.

Over the course of two weeks the SREP was used to test 16 possible SR engine and microphone configurations. Each configuration was evaluated with the same conditions. In every evaluation the same 42 phrase grammar file was used. The test audio consisted of ten speech commands spoken by five users played along with six different background noises played at three different levels of SNR. Therefore each configuration was evaluated on how well it recognized 900 speech command and background noise combinations. A list containing the speech commands, background noises and the SNR they were played at for each configuration is listed in Table 1.

SPEECH COMMAND	BACKGROUND NOISE DESCRIPTION <i>(Each background noise was played once with each speech command at a SNR of 0, 3 and 10 dB)</i>
MAIN SCREEN	Air-conditioned/all windows up/stationary
RADAR	Road noise/driver and passenger window down/30 mph
RECORDS	Engine noise/all windows up/0 – 65 mph
LICENSE STATE	Yelp siren/passenger window down/stationary
10 21 FLORIDA	Wail siren/driver and passenger window down/stationary
VIDEO	Piercer siren/driver window down/stationary
SIRENS OFF	
MALE	
B BOSTON	
MASSACHUSETTS	

Table 1 List of Speech Commands and Background Noises

All background noise sounds were recorded while on patrol with a New Hampshire State Trooper in a State Police cruiser. Speech commands were recorded in the laboratory from five native English speakers. Four of the speakers were male and one was female. The uneven balance between male and female speakers models the fact that the target audience (the police industry) is dominated by males.

The SREP ran inside the Project54 Labcar. The Project54 Labcar is the mid-section of a police cruiser located in the Project54 laboratory. It was used to simulate the acoustic properties the Project54 system will perform under. A picture of the test setup is shown in Figure 3.



Figure 3 Physical Test Set-up

RESULTS

Once all the evaluations were complete, the accuracy of each configuration was compared. The results were compared to determine the most accurate microphone and SR Engine on average as well as the absolute most accurate configuration. The accuracy of each configuration, sorted by microphone, is shown in Figure 4.

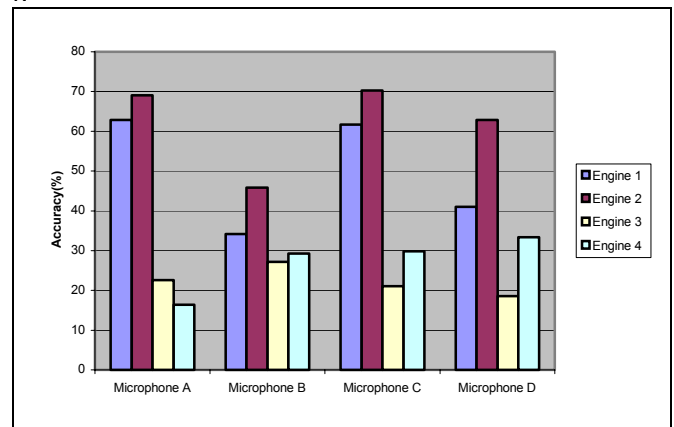


Figure 4 Accuracy Sorted by Microphone

The average accuracy of the four configurations containing each microphone is shown in Table 2. On average, microphone C was the most accurate and microphone B was the least accurate.

MICROPHONE	AVG. ACCURACY (%)
A	42.8
B	34.1
C	45.7
D	39.0

Table 2 Average Accuracy of Each Microphone

The accuracy of each configuration, sorted by SR Engine, is shown in Figure 5.

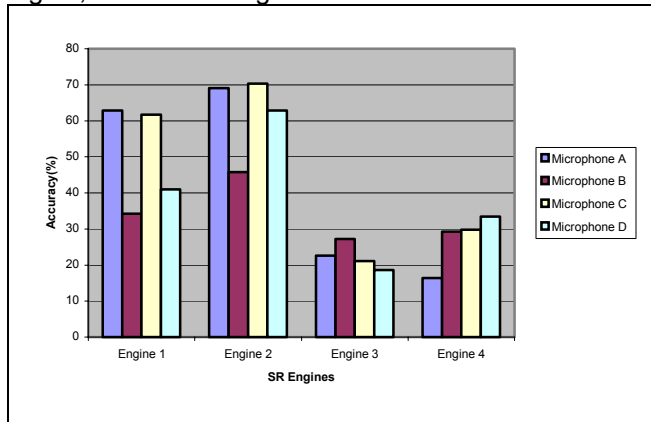


Figure 5 Accuracy Sorted by SR Engine

The average accuracy of the four configurations containing each SR engine is shown in Table 3. On average, SR engine 2 was the most accurate and SR engine 3 was the least accurate.

SR ENGINE	AVG. ACCURACY (%)
1	50.0
2	62.0
3	22.4
4	27.2

Table 3 Average Accuracy of Each SR Engine

On average the most accurate microphone was microphone C and the most accurate SR engine was engine 2. Also, when combined, the microphone C and engine 2 configuration was the most accurate configuration. A plot of the accuracy of all 16 configurations is shown in Figure 6.

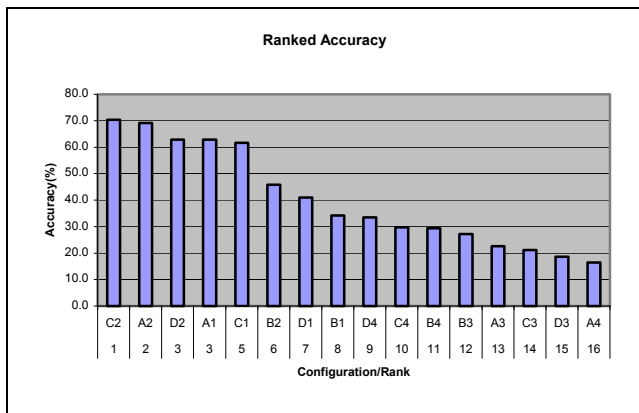


Figure 6 Accuracies Sorted by Rank

Accuracy levels varied dramatically from one configuration to the next. The highest accuracy was 70.1 percent that was achieved by configuration C2. The lowest accuracy was 16.4 percent, which was achieved by configuration A4. Accuracy results from the SREP were more dependant on which SR engine was used rather than on which microphone was used. The eight most accurate configurations all contained either engine 1 or engine 2. The eight least accurate

configurations all contained either engine 3 or engine 4. On the other hand, the division of configurations with a particular microphone was not as clear-cut.

As previously mentioned, each configuration was tested at three different signal-to-noise ratios. A plot showing each configuration's accuracy at each of the three signal-to-noise ratios is shown in Figure 7. As shown in Figure 7, configuration C2 was one of the top performers at higher signal-to-noise ratios of twenty and seven dB and it was clearly the most accurate when the signal-to-noise ratio was at its low point of zero dB.

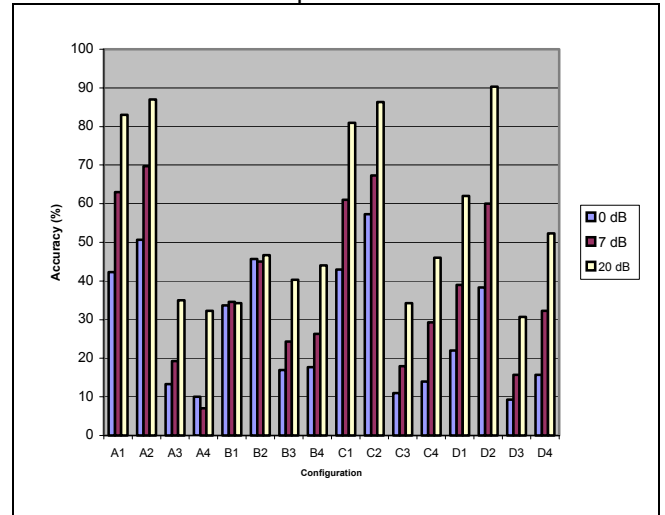


Figure 7 Accuracy at three signal-to-noise ratios

In order to determine which SR engine/microphone configuration performed most efficiently when put through the SREP, a scoring system was developed. The scoring system took into account the fact that falsely recognized commands were worse than unrecognized commands for the Project54 system. When a command is unrecognized in the Project54 system, the system simply remains in the same state it was in before the command and the user needs to repeat the command. If the command is falsely recognized then the system performs that command instead of the desired command. The user then has to possibly undo the falsely recognized command to return to the previous state and restate the command. The scoring system will allow a more precise prediction of which configuration will perform most efficiently as part of the Project54 system, than accuracy alone. To demonstrate the effect of false recognitions on user productivity let us look at the performance of the system over a span of ten commands uttered by a police officer using the Project54 system. If all ten commands are correctly recognized than all ten commands were useful. On the other hand let us say the SR engine misheard one of these commands. If the engine could not recognize the command the officer would have to repeat the command and thus only nine out the ten commands would be useful commands. However if the engine falsely recognized the command and it is a command that needs to be undone, which about half of them are, then the officer would need to undo the falsely recognized

command and restate the desired command. In this case only eight out of the ten commands would be useful commands.

The efficiency scores were determined by awarding 1.5 points for a correctly recognized phrase, 0.5 points for an unrecognized phrase and zero points for falsely recognized phrase. These point values were determined from the ten phrase example earlier in this section. For every unrecognized command, one command is wasted. For the average falsely recognized command, 1.5 commands are wasted. The scores were scaled to have a maximum of 100 and a minimum of zero.

Configuration C2 had achieved the highest efficiency score of 72.4 closely followed by A2, A1 and C1 that were all within one point of C2. Configuration A4 had the lowest score, which was 43.7. A plot of all the efficiency scores is shown in Figure 9.

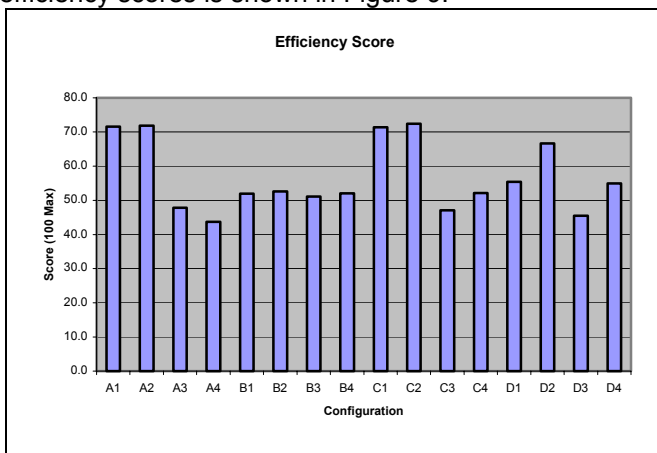


Figure 9 Efficiency Scores

DISCUSSION

The microphone and SR Engine combination that achieved the highest efficiency score, as well as the highest accuracy, when evaluated by the SREP, was configuration C2. Since C2 is the configuration that performed best in both accuracy and efficiency it is the configuration that is recommended for use with the Project54 system.

The engines' accuracy measurements may seem less than acceptable for a commercial application with the highest accuracy achieved being 70.1 percent. However, much higher accuracies are expected when the system is used in the field. There were a number of factors that caused the accuracies to be so low. The first factor was that five people that recorded the speech commands had very little or no experience with speech recognizers or the Project54 system. The second factor was the limited amount of training for the speaker dependent engines and the use of default settings for all the engines. Another factor was that the speakers that played the voice and background sounds were placed together in front of the microphone, so the background

noise was coming straight at the microphone which poses a problem for directional microphones and array microphones that are designed to filter out noise that comes from the side and focus on the sound directly in front of the microphone. Also, although the SNR was set when the audio left the speakers, the SNR was actually lower when it reached the microphone due to various noises in the testing environment.

Due to the reasons stated above it is expected that each configuration will be more accurate and efficient in the actual police cruiser environment than during tests with the SREP. This was known before the configurations were evaluated using the SREP. The goal of the SREP was not to find a microphone and speech engine combination that achieved a desired absolute accuracy or efficiency score. Rather the goal was to determine which of the configurations was more accurate and efficient than the other configuration tested.

CONCLUSION

After evaluating the performance of each of the 16 possible microphone and speech recognition engine combinations under identical conditions in a simulated police cruiser environment, one combination was clearly the most accurate and efficient. The combination of speech recognition components have since been tested in five New Hampshire State Police cruisers and will soon be installed in 300 police cruisers that are being outfitted with the Project54 system

REFERENCES

- [1] <http://www.catlab.unh.edu/>
- [2] Power K J. "The Listening Telephone – Automating Speech Recognition over the PSTN," *BT Technical Journal*, Jan 1996, Vol 14, No. 1, pp112-126.
- [3] R. D. Johnston. "Are Speech Recognisers Still 98% Accurate, or has The Time Come to Repeal 'Hyde's Law',?" *BT Technical Journal*, Jan 1996, Vol 14, No. 1, pp165-176.
- [4] H. Agaiby, C. Fyfe, S. McGlinchey, T. J. Moir "Commercial Speech Recognisers Performance Under Adverse Conditions, A Survey," ESCA NAID Workshop, April 1997.
- [5] Dale Rogerson, *Inside COM*, Microsoft Press, Redmond, WA, 1997
- [6] A. L. Kun, W. T. Miller, W. H. Lenharth, "Modular System Architecture for Electronic Device Integration in Police Cruisers," *Proceedings of the 2002 IEEE Intelligent Vehicle Symposium*, Versailles, France, June 18-20, 2002
- [7] B. J. Vinciguerra, "A Comparison of Commercial Speech Recognition Components for use with the Project54 System" MS Thesis, University of New Hampshire, Durham, 2002