April 10, 2004

Lakshman One
School of Engineering Science
Simon Fraser University
Burnaby, British Columbia
V5A 1S6

Re: ENSC 440 Post-Mortem – Voice Recognition System in MP3 Players

Dear Mr. One:

The attached document, *Post-Mortem for a voice recognition system in MP3 players,* summarizes the overall development process of our ENSC 440 project. We've worked with Start Labs Inc. on controlling their MP3 players via voice commands. Our project was to design the voice recognition module of the MP3 player. We feel confident that our design has met Start Labs Inc.'s needs and expectations.

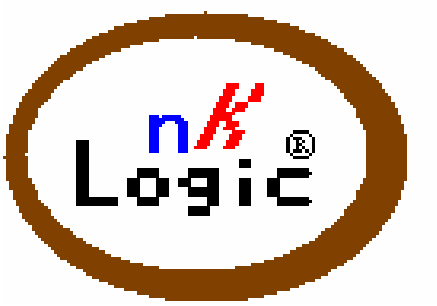
This document illustrates the nature and current state of the project, and problems encountered during the test phase. It also discusses project management, future plans, and personal experience in the participation of the project.

The nK Logic Group consists of two experienced senior engineering students: Won Kang and Garet Kim. We look forward to your feedback and suggestions. Please feel free to contact me by phone at (604) 785-5933 or by e-mail at gkim@sfu.ca. Thank you for you attention.

Sincerely,

Garet Kim
nK Logic

Enclosure: *Post-Mortem for a voice recognition system in MP3 players*



Post-Mortem for a

Voice Recognition System in MP3 Players

Project Team: Won Kang
Garet Kim

Contact Person: Garet Kim
gkim@sfu.ca

Submitted to: Lakshman One – ENSC 440
Nakul Verma – ENSC 440
Mike Sjoerdsma – ENSC 305
School of Engineering Science
Simon Fraser University

Issued date: April 10, 2004

Table of Contents

1. Introduction

Start Lab Inc. is a company specializing in wireless MP3 players. Its MP3 player utilizes Bluetooth for connection between the main unit and the remote controller. It also features voice-activated commands which give the users some degree of autonomy. The nK Logic has participated in this project and was responsible for the research and development of this voice activated controller unit.

By the end of this semester, we, the nK Logic, have finished developing the prototype. Although the basic operation was acceptable, the most important factor, the success ratio did not quite reach the satisfactory level. So we have put much time and effort in optimizing and testing the final prototype device. At the same time, we found alternative solutions using different technology or a new IC.

2. Current State of the System

The voice recognition module, developed by the nK Logic Group, allows the user to control the MP3 player by speaking aloud into the microphone. Since the Start Labs' MP3 decoder unit is still under development, the Voice Controlling Unit (VCU) simply outputs messages to the host computer that identifies which voice control command has been recognized.
The voice recognition can be viewed as a series of sequential processes. This is illustrated in Figure 1.

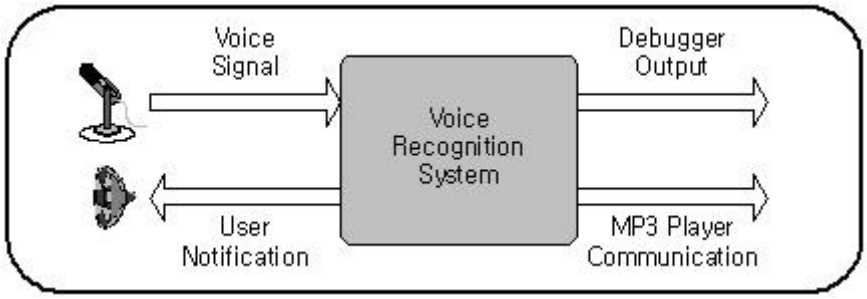


Figure 1: Configuration of Voice Recognition Process

2.1 Speech Technologies

In a VCU, accuracy of the recognition measures integrity of the system. To improve it, Voice Extreme™ supports various speech technologies, such as Speech Synthesis (SS), Speaker Independent (SI), Speaker Dependent (SD), Continuous Listening (CL), WordSpot (WS), and Speaker Verification (SV). For our application, both SD and CL technologies are employed to achieve the two-level grammar structure with high success ratio. The WS was also used in order to improve the reliability of the system later on. The summary of the technologies is shown in Table 1.

Table 1: Speech Recognition Technologies

Technology	Detail
SD: Speaker Dependent	• Words must be recorded prior to recognition • Training required • Higher success ratio over SI
SI: Speaker Independent	• Weights are generated from the utterance. • Training not required • More Convenient, but expensive
CL: Continuous Listening	• Continuously repeats the SI/SD process • Does no mean detecting continuous speech
WS: WordSpot	• Continuously looks for a trigger word • Does not require pause, but there can be only one trigger word

2.2 System Operations

A Speaker Dependent VCU requires three basic programming modules: Training, recognition, and debugging modules. First, the training module prompts the user to say each command twice. When those two commands sound similar, the module averages them, and saves the average into the memory. In the recognition module, it receives the command input and compares it with saved commands to find a match. Once the recognition module finds a match, it tells to the debugging module which command has been received and recognized (Debugging module is to be explained in the later sections). Figure 4 illustrates the operation of this system in a flow chart.

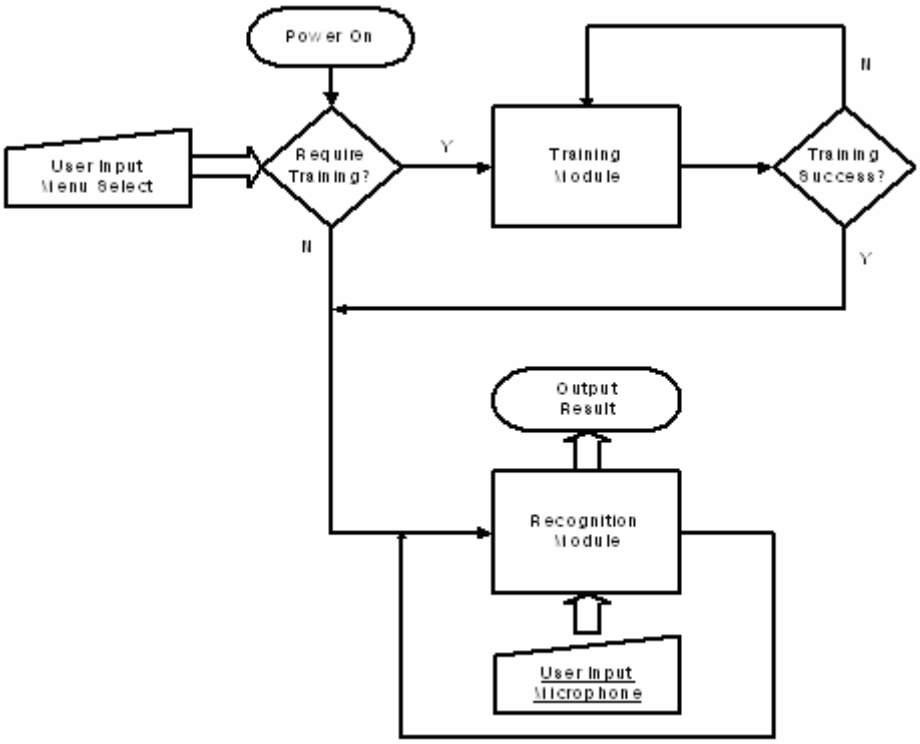


Figure 2: Overview of a voice recognition system

2.3 User Input

The VCU receives signals from two sources: Push buttons on the board and the microphone. Push buttons are used for system settings and interrupts while the microphone accommodates the user's voice commands.

2.4 Debugger

A debugger is an important tool for monitoring the overall operation of the system. In this system, the host computer is connected to the board through RS232. Once the development board acknowledges the voice commands, it tells the host computer (debugger) which voice command has been accepted. Then, the debugger outputs the corresponding messages.

Although a text-based debugger was suggested at the proposal stage, the nK Logic Group has decided to create a GUI version of the debugger to enhance the development and testing processes. Since the VCU will be a part of the Start Labs' MP3 player, the

nKLogic designed a debugger that imitates an mp3 player display using Microsoft Visual C language. At this point, we felt that it was still insufficient to prove actual interaction with MP3 players. Therefore, we decided to control the Winamp®, a popular MP3 encoding program in Windows, through the debugger. Figure 3 shows the GUI debugger and the Winamp window.

Figure 3: Debugger and Winamp MP3 Player

3. Problems Encountered

3.1 Problems

Once the development of the Voice Extreme was done, we started testing the performance of the system. The following is the summary of the test on Voice Extreme:

- Success ratio close 100% with playing pre-recorded commands
- Success ratio only around 60% with live commands
- Ratio drops as the number of commands increases
- Extremely sensitive to the background noise, and this is likely to cause false detections
- Not much flexibility (e.g. sensitivity) in programming to enhance the success ratio

Although this system was developed as a prototype, we discovered that its overall performance was too poor to be implemented into any kind of real systems. As such, we decided to look for alternative solutions. We found two possible solutions: Upgrade on Voice Extreme using WordSpot technology and a new chip, Sensory Inc.'s RSX-4 Series.

3.2 Alternative Solution 1: Voice Extreme with WordSpot

The bigger problem with our system was high false detection ratio. For instance, when people chatted casually in the background, the system frequently recognized some commands and executed them although there were no command words spoken. In order to distinguish between casual chatting and actual commands, we employed WordSpot technology. As explained in previous sections, WordSpot recognizes the command (the trigger word) even when it is spoken continuously, without any pause before and after. We modified the software in such a way that the trigger word had to be recognized before the normal recognition operation starts.

In order to have high success ratio, we lowered threshold level of recognition with a distinctive, unusual trigger word, such as "iRiver". The test showed that this has significantly improved the reliability of the system.

3.3 Alternative Solution 2: RSX-4 Series

Another solution was to select another chip and program. We found this RSX-4 series from Sensory Inc. The functionalities are similar to Voice Extreme. However, this has enhanced WS (4 to 10 trigger words in parallel) and speech synthesis quality. RSX-4's software tools are quite expensive, ranging from US $3,000 to $5,000.

When we ordered and received this board, we only had two weeks to work with this IC. So we were only able to do some preliminary testing comparing performances of two ICs: Voice Extreme and RSX-4.

3.3.1 Background Noise

We first recorded a word using Cool Edit Pro, an audio program and added different amount of white noise to observe corresponding success ratio. All the success ratios were calculated based on 50 tries, and white noise magnitude of 2 specified in Cool Edit Pro was referenced as 100% on the horizontal axis. The result is shown in Figure 4.

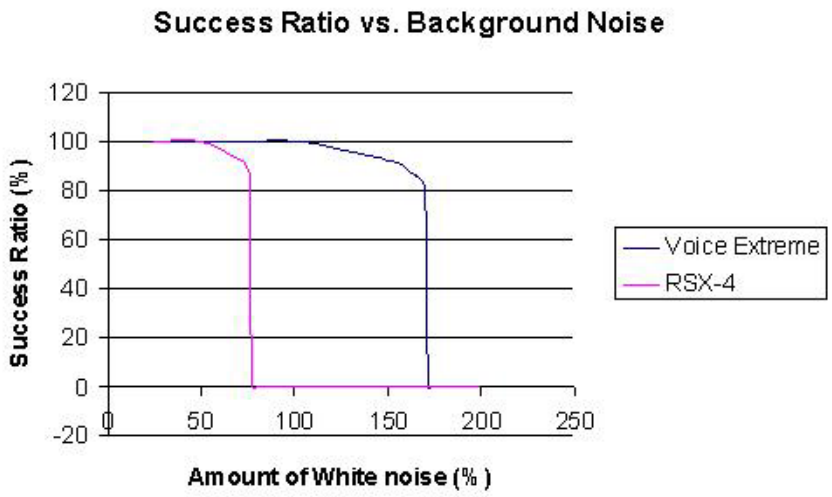


Figure 4: Success Ratio vs. Background Noise

This result implies that the Voice Extreme responds more reluctantly to added noise. This explains why there are so many false detections when testing with Voice Extreme.

3.3.2 Rate of the Speech

To see how systems respond to different speed of the speech, we changed the rate of the speech in Cool Edit Pro and measured the success ratio, as explained in the previous section.

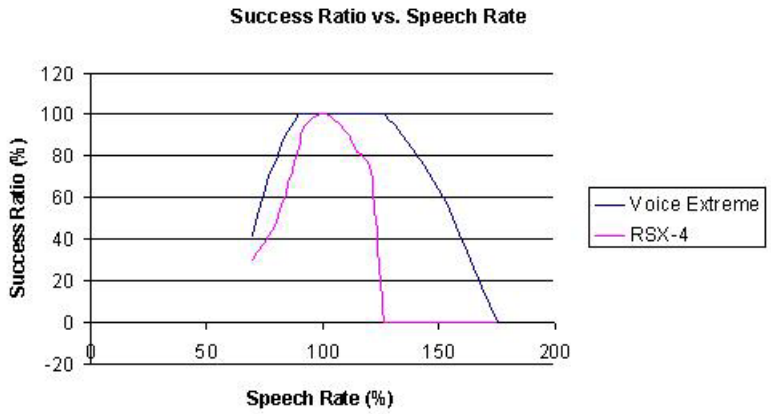


Figure 5: Success Ratio vs. Speech Rate

The result illustrates that the Voice Extreme can achieve high success ratio over broader rage of speech rate than RSX-4 can do. This again demonstrates that the Voice Extreme recognize the command rather too easily, which causes wrong detections to take place when there are many commands to be compared.

3.3.3 Volume of the Speech

We also adjusted the volume of the speech and compared the success ratio. Figure 6 shows the result.

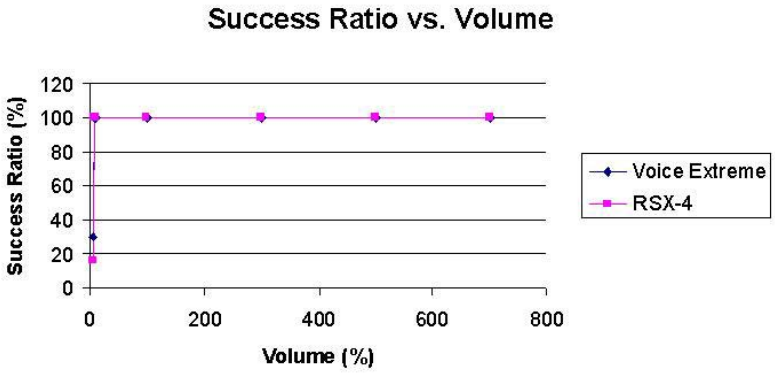


Figure 6: Success Ratio vs. Volume

Figure 7 specifically shows the results around 6% volume of the original recording.

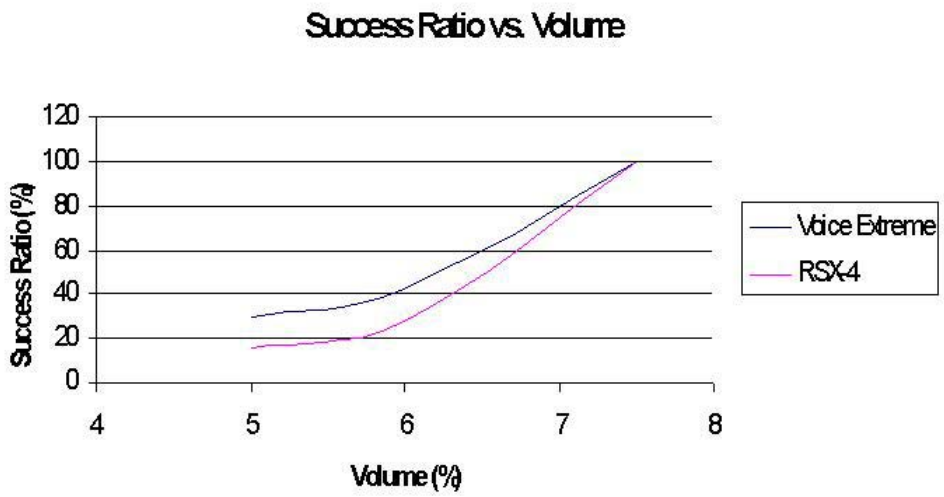


Figure 7: Success Ratio vs. Volume Near 6%

Figure 7 shows that the Voice Extreme recognizes more than RSX-4. This again proves the Voice Extreme is much more receptive in recognition.

3.4 Conclusion

Voice Extreme has higher success ratio over broader range of noise, rate and volume of the speech. This may be perceived as a good characteristic. However, this actually implies that Voice Extreme is too receptive and too responsive when there is more than one command to be looked for. This is why there is too much false detection when tested. In an attempt to reduce such false detections, WordSpot technology can be implemented on the existing system. Yet another solution would be using RSX-4 IC, which is more reliable and gives more flexibility in development.

4. Budgetary and Time Constraints

4.1 Budget

The following table summarizes the total expenditures for this project.

Table 2: Budget Comparison

Item	Estimated Cost ($)	Actual Cost ($)
Voice Recognition Toolkit	170.00	318.00
Development Software	50.00	Included
Acoustic accessories	20.00	Included
Cables	20.00	Included
Case	20.00	Included
Contingencies (15%)	40.00	0
Total	320.00	318.00

Start Lab Inc. agreed to purchase all necessary parts for this project. As a result, nK Logic loaned voice recognition development kit, which includes the board, software, cable, and case from Start Lab. Thus, nK Logic did not need to create any expenditure for this project.

4.2 Time

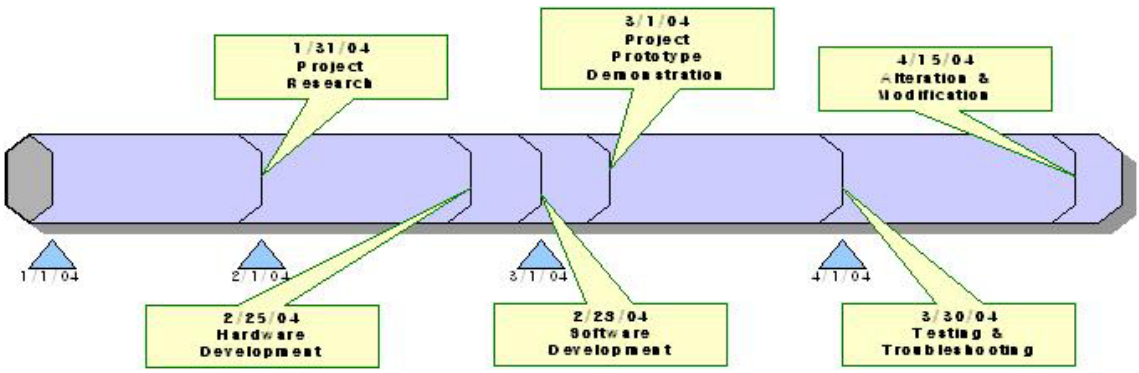


Figure 8: Estimated Schedule

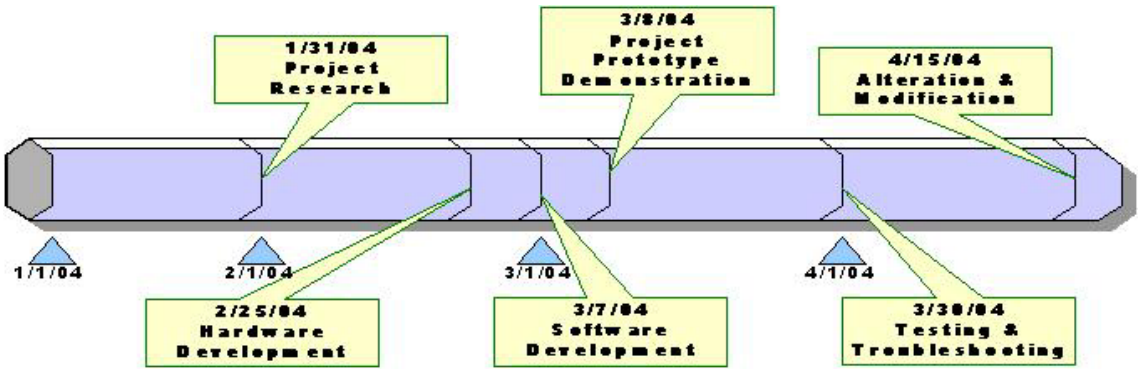


Figure 9: Actual Schedule

We are proud that wee have successfully managed the project time constraints. The software development had to be delayed because the shipping took a few more days than we had expected.

After we selected the ICs, we studied the programming materials and started writing the code using the simulator before the development board was delivered. This allowed us to have enough time for the testing and optimization. The majority of our time was spent in performing the system testing to see if the board was actually applicable for real world application, and this was an obligation to delay the demonstration by a week.

5. Future Plans

As was aimed, the prototype voice recognition system has been developed, and we are now looking forward to see the whole system's being integrated into a marketable device. In order to be used in the actual device, our voice recognition system needs to be modified appropriately. We will have to build a hardware circuit to connect it to the MP3 player unit. It will also require PCB designs, power sources, memory data/address lines and I/Os to be added.

Most of all, we would like to put more effort in ameliorating the quality of the system. The ultimate goal would be to build a system on a DSP and program the software using more enhanced algorithms. This way, we will have more control over the system and we expect that the performance will be much superior to that of the current system.

We have described the different characteristics of two specific ICs and the related technologies. More detailed comparisons including different ICs and DSPs would provide Start Lab Inc., a means to choose the best solution for their application.

6. Personal Experience

We believe that our team worked very efficiently in terms of organizing and managing the project. Each group member took full responsibilities for the part they were assigned. This made the integration of the system much easier and faster. Also, we learned that a well-defined specification is a very efficient tool for project management. Following the functional and design specifications helped us to develop the system on time without any major difficulties.

Throughout the project, each member has learned both technical and social skills in various ways.

Won Kang

First of all, I learned the importance of team work from this course. The project was expected to be too much for only two members. However, assigning the appropriate work to each member from the beginning helped us to work efficiently. For example, each of us was in charge of the areas that each has some experience with. In addition, I realized the importance of documentations. The functional specification and the design specification were the basis for our project and saved us a lot of time in design specific decision-makings.

I also realized that the skills and experience that I gained from previous engineering courses were actually applicable in real projects. The theories and practices from the classroom were invaluable to the success of this project, from programming a custom IC to testing it.

Having finished this project, I feel a lot more confident in facing new challenges. This project has taught me that problem-solving is not about how to create the solution, but is about how find applicable concepts or ideas and apply them.

Garet Kim

For the last thirteen weeks, I participated in assembly/C programming and quality testing of the system. At the beginning, we weren't sure if we would finish this project on time since our group consisted of only two members. Now having finished the project, I realize the power of documentation. A well-defined project schedule and well-written specifications were the best resources which allowed us to keep track of our progress and the system design throughout the whole semester. Gantt charts were especially useful in case of multitasking.

The most challenging part of this project was that we didn't have much control over the manufacturer's voice recognition algorithm. Sensory Inc. allowed programmers to vary only limited factors, and this made it difficult for us to enhance the system performance. Voice recognition algorithms are mathematically well-established concepts. In the future, I'd like to be involved in implementing voice recognition chip that can handle advanced voice recognition algorithms.

7. References

1) Sensory Inc. http://www.sensoryinc.com
2) Phyton http://www.phyton.com/htdocs/tools_se/main.shtml?tools_se.shtml~D