Investigation and Testing of Text-to-Voice Software for an Automated Dictation Device

Interim Report CPSC 503

Brian David Fox
foxbd@cpsc.ucalgary.ca

December 4, 2005

Abstract

This study compares and tests available commercial and open source voice-to-text software for use with the Alan project. After a comparison of seven applications was completed, the results yielded two candidates for further evaluation and testing, Microsoft XP Voice Recognition (VR) and Dragon Naturally Speaking. The tests involved reading in text from O'Reilly's Learning UML with a microphone as well as a prerecorded wave file into both applications. The results for the live text were an accuracy rate of 88% for both applications for the live read text. This meant that available features that each application had would be the determining factor in terms of which application was the better choice for the Alan project. Dragon Naturally Speaking had a transcribe feature that allowed prerecording speech and processing it with the application at a later date. The XP VR software did not have this feature. This was one of the important criteria required for the Allan Project and hence made Dragon a better candidate for immediate use. Further experiments were also conducted on Dragon using a larger dataset of text and it was determined that the software's accuracy was 90%.

Introduction

The goal of this investigation was to determine the most effective voice-to-text software in terms of accuracy and features available with currently existing applications. It was necessary to compare both commercial and open source applications to determine possible candidates for testing. Due to the limited time available, it was essential to use a comparison method to minimize the number of applications to be tested. This provided valuable information that could be used to determine which voice-to-text software is most suitable for use with the Alan project. The Alan project being a proposed project aims at recording the events at a software engineering meeting and processing the recording using voice-to-text technology. By determining the most effective software available the Alan project has a higher probability of success.

Other projects currently underway, in respect to voice-to-text, include the University of Colorado's *Sonic Project* that emphasizes speech recognition development[1] as well as the Sphinx4 project that is currently a collaborative effort by Carnegie Mellon University, Sun Microsystems Laboratories, Mitsubishi Electric Research Labs, and Hewlett-Packard's Cambridge Research Lab[2] also emphasizing speech recognition development.

In order to complete this investigation in the allotted time frame, we first compared a number of available applications qualitatively. By comparing the applications features and published accuracy rates, we were able to shortlist the applications that were candidates for actual testing. Using this method reduced the need to test all the applications, which is time consuming and was not possible given the time frame of this investigation.

Methodology

In order to do this study there were a number of components required: comparing the applications, choosing the test candidates, and finally testing the candidates. The comparison component included 7 different applications, some of which were from the same manufacturer but with different available feature sets. Ideally, testing all 7 of the applications would have been the best approach, however time did not permit this approach and as a result the decision was made to test two of the most suitable applications determined by the qualitative comparison.

As mentioned above, features make the difference between possible acceptance and rejection in terms of testing candidates. The investigation used 21 categories for overall comparison of the 7 applications shown in **table 1**.

Application Comparison Table

	Sphinx 4	**Dragon Naturally Speaking Pro Edition**	**Dragon Naturally Speaking Standard Edition**	**IBM ViaVoice Standard Edition**	**IBM ViaVoice Pro USB**	**Sonic**	**Windows XP Voice** Recognition
Version	4.0	8.0	8.0	10.0	10.0	2.0	5.0
Hardware Requirements	-Intel Pentium 4, 1.7 Ghz -900 Mb RAM -Cache 1800 M	-Intel Pentium III processor/500 Mhz or AMD equivalent -256 Mb RAM - Sound Blaster 16 or equivalent 16 bit recording soundcard -CD-ROM -Web account f	-Intel Pentium III processor/500 Mhz or AMD equivalent -256 Mb RAM - Sound Blaster 16 or equivalent 16 bit recording soundcard -CD-ROM -Web account f	-Intel Pentium 266Mhz -256 L cache, -192 Mb RAM -500 MB hard drive space -Windows compatible 16 bit soundcard (Windows 98SE and above) -CD-Rom (Quad speed and above)	-Intel Pentium 266Mhz -256 L cache, -192 Mb RAM -500 MB hard drive space -Windows compatible 16 bit soundcard (Windows 98SE and above) -CD-Rom (Quad speed and above)	-Intel Pentium 4, 2.2 Ghz	-Intel Pentium 450Mhz or better -256 Mb Ram or above -590 Mb hard drive space -Super VGA (800 X 600) or higher
Possible Number of User Accounts	N/A	Technically Unlimited	Technically Unlimited	N/A	N/A	N/A	Technically Unlimited
Cost	Open Source	$851.47 CDN PER USER	$106.96 CDN PER USER	$49.99 CDN	$189.99 CDN	Open Source	Free with Windows XP Professional And XP office Professional.
English as a Second Language Testing	N/A	N/A	N/A	N/A	N/A	YES	N/A
Accuracy Small Dataset 100 Words	Between 77.81% and 98.81% Depending On Platform	Up to 99% (Promotional Website data)	Up to 99% (Promotional Website data)	Up to 90% (Promotional Website data)	Up to 90% (Promotional Website data)	Between 98.8% and 89.1% depending on task.	Up to 90% (Promotional Website data)
Accuracy Medium Dataset 20,000 Words	Between 80.91% and 85.86% Depending on platform	Up to 99% (Promotional Website data)	Up to 99% (Promotional Website data)	Up to 90% (Promotional Website data)	Up to 90% (Promotional Website data)	Between 98.8% and 89.1% Depending on Task.	Up to 90% (Promotional Website data)
Accuracy Large Dataset 64,000 Words	Between 80.91% and 85.86% Depending on platform	Up to 99% (Promotional Website data)	Up to 99% (Promotional Website data)	Up to 90% (Promotional Website data)	Up to 90% (Promotional Website data)	Between 98.8% and 89.1% Depending on Task.	Up to 90% (Promotional Website data)

Microphone Required	N/A	Noise Canceling	Noise Canceling	Noise Canceling	Noise Canceling	N/A	Noise Canceling
Platform	Windows, Linux	Windows, Mac	Windows, Mac	Windows	Windows	Windows, Mac OS X Sun Solaris, Linux	Linux
Compatible Applications	N/A	Most Windows Applications	Most Windows Applications	ViaVoice SpeakPad , Microsoft® Word 2002, 2000 or 97	ViaVoice SpeakPad , Microsoft® Word 2002, 2000 or 97	N/A	Windows Applications
Development Language	Java	C++, Python	C++, Python	C++	C++	ANCI C	C++
Software Requirements	Apache Ant Java 1.4 or Higher Linux or Windows 98 Or Higher	Windows XP (Sp1 or Higher;) Home and Professional, 2000 sp4 or Higher	Windows XP (Sp1 or Higher; Includes XP Tablet Edition) 2000 sp4 Or higher	Microsoft® Windows 98SE, ME or XP Home	For Microsoft ® Windows 98SE*, 2000 Professional (Service Pack 2) And XP Home & Professional Editions	Windows, Mac OS X	Windows XP (Sp1 Or Higher) , XP Office
Company/ Educational Institution	Carnegie Melon University	Scansoft	Scansoft	IBM	IBM	University of Colorado	Microsoft
Prerecorded Data Feature (Transcribe)	The application can handle raw data or cespa files. However there is a way to transcribe .wav files.	The application has transcribe a feature for inputting digitally recording for later processing.	No transcribe Feature.	No transcribe Feature.	The application has transcribe a feature for inputting digitally recording for later processing.	N/A	No transcribe Feature.
Microphone Configuration	N/A	Handheld or Headset	Headset	Headset	Headset	N/A	Headset
Discrete/ Continuous Dictation	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous
Support for Cordless Microphone	N/A	YES	NO	N/A	N/A	N/A	N/A
Speech Recognition Engine	Hidden Markov Model (HMM)	Hidden Markov Model (HMM)	Hidden Markov Model (HMM)	Hidden Markov Model (HMM)	Hidden Markov Model (HMM)	Hidden Markov Model (HMM)	Hidden Markov Model (HMM)
First Release Date	1987	N/A	N/A	N/A	N/A	2001	N/A
Training Time	N/A	1.5 hrs Minutes. Accuracy Increases with More training	1.5 hrs Minutes. Accuracy Increases with More training	10 to 30 Minutes. Accuracy Increases with More training	10 to 30 Minutes. Accuracy Increases with More training	N/A	1 or More Hours

Table 1

Once the information from the Application Comparison Table was reviewed, two applications were chosen for further evaluation and testing. Testing entailed using a standard headset for dictation in a controlled environment. The controlled environment is a quiet room with no auditable exterior noise. It is important to note that the microphone was a non-noise-canceling microphone. The reason for this was to emulate, as closely as possible, the microphone quality of the robot being used in the Allan project.

The text dictated into the applications was read in as well as fed in using a prerecorded file known as the transcribe feature of the Dragon Naturally Speaking application. The

written text that was used for dictation was an excerpt from O'Reilly's Learning UML which contains content words that might be used in software engineering meetings.

Results

Subsequent to reviewing the information in the table, the decision was made to test Microsoft's XP Voice Recognition system and Scansoft's Dragon Naturally Speaking software. This initial decision was based on the accuracy results row in table 1 as well as cost and feature evaluations.

In order to test the applications, the following metrics were used to compare the read in or prerecorded text, original correct text, with the output of each of the applications; *MW*, *AW*, and *WW* each constituted a single error.

- **MW:** Missed Word: Words that were spoken but missed in the transcript.

- **AW:** Additional Word: In this case the application adds more than one word when only one word is spoken. The metric used in this case was:
(additional words) –1 = error words.
The reason for this is that the application was legitimately trying to produce the said word.

- **WW:** Wrong Word: In the case that the wrong word appears in the transcript but it was subjectively very close to the intended word a 1c will accompany the count otherwise it will be marked with 1. For example if the word spoken was UML and the application produced UNL than this constitutes a 1c not a 1.

The following tables show the error results generated by the testing process:

Word Error Rate (WER) Per Dictation Session: Original Generated and Prerecorded Text Results (real time and prerecorded generation) Short Dataset

Application	MW	AW	WW	Total Errors	Accuracy
Dragon (real time dictation)	1	13	11 (including nine 1c errors)	25	88%
Dragon (Prerecorded)	1	13	11 (including nine 1c errors)	25	88%
Microsoft XP VR (real time dictation)	1	13	11 (including nine 1c errors)	25	88%
Microsoft XP VR (prerecorded)	N/A	N/A	N/A	N/A	N/A

Table 2

Please note: the metric for determining the accuracy rate was as follows:
(total original words -total errors)/(total original words). For example XP's accuracy was measured as follows (209 – 25)/(209) = .8803 yielding an 88.03% accuracy.

As can be seen in the above tables, both of the test applications performed identically on live dictation in terms of accuracy. However, when the original prerecorded wave file was transcribed into Dragon Naturally Speaking, the accuracy fell from 88% to 81%, a difference of 7%.After re-examining the wave file used it was determined that speed at which the text was spoken was rushed in places and that this was the cause of the accuracy degeneration. This shoed that speech speed had a significant impact on accuracy. Retesting with a new wave file generated the same accuracy as the read in speech, 88%. Because Microsoft's XP VR dose not have a transcribe feature, the wav file that was previously used with Dragon was transformed into an MP3 file and fed directly into the microphone jack in the rear of the computer. This yielded no results at all using the original wave file. After retesting with the new wave file the XP VR did generate some text, however, the content was greatly degraded in accuracy. The XP VR generated only 97 words of gibberish making it impossible to apply the above metrics to it. It could not be ascertained at this time why the XP VR performed so poorly on the prerecorded text.

It is important to note that above testing was done on a short text data set. For the purpose of these experiments, *Short Text* refers to dictation of approximately 200 words were as *Long Text* refers to dictation of 1000 or more words. These metrics are valuable for comparison of accuracy rates of testing. Because the XP VR did not generate any usable test results using the prerecorded text on the short dataset, and given the time constraints of the experiments, only Dragon was used to test the long dataset. The results were as follows.

Dragon Word Comparisons Long Dataset

O'Reilly Original Text		**Dragon Prerecorded Text Output**	
Words	1,210	Words	1,221
Characters No Spaces	6,615	Characters No Spaces	6,595

Table 3

Word Error Rate (WER) Prerecorded

	MW	**AW**	**WW**
Dragon	18 words	27 words	66 words including 22 1c words
Total Errors	111		
Accuracy	90%		

Table 4

As can be seen in the above table Dragon Naturally Speaking performed better on the long dataset of text, an increase in 2% using the longer dataset.

Another observation we gained from testing was that the accuracy rate increased dramatically after multiple training sessions were completed with each application. Training consisted of reading recommended text into each application in order to train it to recognize the individual's voice. Microsoft's VR needed four 7 to 9 minute training session and Dragon required three sessions. Two of these recommended texts were 7 to 9 minutes and the third text was 1 hour and 10 minutes. To ensure that the retesting was unbiased using the new wave file, an additional 6 training sessions were added to the XP VR application ranging from 7 to 12 minutes in length.

Below are the word count results generated during short dataset testing after training had been completed on both applications.

Word Count Comparison Table Short Dataset

Source	Words	Characters No Spaces
O'Reilly Original Text	209	1,254
Dragon Live Dictation	221	1,283
Dragon Prerecorded Dictation	232	1,233
Microsoft VR Live Dictation	231	1,283
Microsoft Prerecorded Dictation	97	463

Table 5

As a result of both applications having identical accuracy rates regarding live dictation; at no time was the level of accuracy attained that was indicated possible in Dragon's White paper.[3] Given these results, the decision of which application was left to the features available with each application. Because the project requires a transcribe feature that allows for prerecorded data to be inserted into the program and given the test prerecorded text results for the XP VT, Dragon Naturally Speaking was the reasonable choice for immediate use. However, it is possible to use XP's .Dlls library to produce a similar feature.

Discussion

Overall, the approach used in testing the two applications showed that in a controlled environment that the applications had an 88% acceptable rate of accuracy for the short dataset and Dragon yielded 90% accuracy for the long dataset. It is important to note that the tests were conducted using a Native English speaker. To further investigate the accuracy of the software, for use with the Alan project, an English as a second language participant should be used in the testing process. There is information that points to a fall in accuracy when using the Hidden Markov Model, which is the speech recognition model currently used by Dragon and the XP VR.[4]

Another possible avenue to investigate would be to use data from each of the members of the software development team that attend the meeting and apply that dataset to Dragon to ensure thorough testing. This dataset would be recorded at the meeting in the

appropriate environment and hence yield more accurate test results than the previous testing done in a controlled quiet environment.

Although the XP VR did not yield satisfactory test results on the prerecorded text, it is possible that it was a hardware error and not an application error and hardware error. In other words, the Dragon application did not need a hardware device to handle prerecorded text and the XP VR did. The hardware device may be the reason for poor performance. Further testing on another computer is required to eliminate the possibility of hardware error on the original testing computer.

Conclusion

In conclusion, this study was to generate a speech recognition software candidate for use with the Alan project. The ~88% accuracy rate shown by both the XP VR and Dragon live dictation and the 90% accuracy shown by Dragon prerecorded text may be enough to meet the requirements of the Alan project testing process. However, until the accuracy of the summarizing software that is to be used to summarize the text generated by Dragon is known, there is no way of being sure that this accuracy is sufficient for the project. Also testing participants that have English as a second language will further determine if the accuracy of Dragon is a suitable application for the Allan project. Future work in regards to testing this software includes testing the recordings generated at the meetings with both applications. The transcribe feature would have to be implemented using the XP .Dll files.

References

[1]Umit H. Yapanel and John H. L. Hanson. A New Perspective on Feature Extraction for Robust In-Vehicle Speech Recognition. In *Eurospeech Geneva.* Pages 1281 -1284, 2003

[2] Paul Lamere, Philip Kwok, William Walker, Evandro Gouva, Rita Singh, Bhiksha Raj and Peter Wolf. Design of the CMU Sphinx Decoder. In *Erospeech Geneva* 2003

[3] Robert Zick, MD and Jon Olson, MD. Voice Recognition Software Versus a Traditional Transcription Service for Physician Charting in the ED. In *American Journal of Emergency Medicine Volume 19 Number 4* July 2001. Pages 295-298

[4] Ayako Ikeno, Bryan Pellom, Dan Cer, Ashley Thornton, Jason Brenier, Dan Jurafsky, Wayne Ward, and William Byrne, Recognition of Spanish-Accented Spontaneous English. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, Japan*, April, 2003