Models of Throughput Rates for Dictation Revisited: Consideration of Speech Processing Speed

TR 29.3599
November 18, 2002

Patrick M. Commarford
James R. Lewis

IBM Pervasive Computing

Boca Raton, Florida

Abstract

Personal Digital Assistants (PDAs) allow users to perform a multitude of tasks and access an abundance of data. The most prominent drawback of these handheld computers is the slow speed at which information can be input. Commarford and Lewis (2002) modeled throughput speed for a hypothetical dictation input method for PDAs that processes speech at real time. The current modeling effort estimates expected throughput for a PDA dictation input method, assuming likely speeds of speech processing given current device limitations. The models suggest that speech dictation would be competitive with current alternative data input methods, even if speech processing were relatively slow.

ITIRC Keywords

Dictation
Graffiti
Handheld
Handwriting
Modeling
PDA
Soft Keyboard
Virtual Keyboard
Speech
Voice Spelling

Contents

Introduction

Personal digital assistants (PDAs) continue to rise in popularity in the United States and elsewhere. This comes with increased functionality and with the desire to stay connected while away from the home or office. A major disadvantage of PDA use is the extremely slow data input speed. The most representative studies show that users can input text at a rate of 12.62 words per minute (WPM) with a 4% error rate using a soft virtual keyboard on a PDA (Zha & Sears, 2001) and can input approximately 4.95 WPM (5% error rate) with the Graffiti[1] handwriting recognition system (Sears & Arora, 2001). Assuming minimal time to correct errors, we can conclude that users can input no more than 12 corrected words per minute (CWPM) with the input technologies currently available for stand-alone PDAs (i.e., not docked or otherwise connected to a personal computer). This pales in comparison to rates that users can achieve with the keyboards of personal computers or laptops.

To determine how much more proficient, if any, a user could expect to be with a speech dictation input method for a PDA, Commarford and Lewis (2002) modeled throughput (defined as CWPM) for a range of speech recognition accuracies and a range of user correction speeds at two different speaking rates (100 WPM and 150 WPM). They compared the expected throughput rates to the 12 CWPM rate associated with soft keyboard input. This investigation, however, assumed a system that processed speech at real time. Although this performance is widely available with desktop dictation software, initial versions of similar software for PDAs will likely not achieve such performance. Given the extremely large benefit of dictation hypothesized by the Commarford and Lewis models (potentially twice as fast as soft keyboard input), it is reasonable to consider whether PDA dictation applications that are not able to achieve real time speech processing would also be beneficial.

This report describes performance modeling for a hypothetical PDA dictation input method running at speeds ranging from real time to 4X slower than real time, given various system recognition accuracies and a range of correction speeds for the 100-WPM speaker and the 150-WPM speaker.

[1] Graffiti is a registered trademark of Palm Computing, Inc.

Terms and Definitions

Processing Speed (running at nX). Defined in terms of real time; it takes n minutes to process 1 min of speech. For example, 2X means that it takes 2 min to process 1 min of speech. We modeled two rates of speech, 100 WPM and 150 WPM. When nX is used to model throughput for the 100-WPM speaker, nX means that the system can process 100 words in n minutes. When nX is used to model throughput for the 150-WPM speaker, nX means that the system can process 150 words in n minutes.

The models assume that once the engine gets to a point at which it has not processed anything spoken in the previous 60 seconds, it will quit accepting dictation and users will have to wait for the system to catch up. During this time, the system will process speech at a faster speed of 2/3nX.

Recognition Accuracy. Percentage of words the system correctly recognizes. We modeled the range of system recognition accuracies from 50%--100%.

Correction Speed. The mean time it takes a user to correct a misrecognition.

No such dictation software for a PDA currently exists, but we can estimate that typical correction speeds for experienced PDA users would be 15-20 seconds. This estimate is based on the faster speed (13.2 sec/correction) that Lewis (1999) observed with the desktop product. We modeled correction times ranging from 10-25 sec.

Speech Rate. How fast the user speaks (in terms of words per minute; WPM).

Throughput Rate. How many CWPM the user is able to enter.

Text. These models consider text that contains no capitalization except for characters that follow punctuation marks.

The Models

Recognition Accuracy by Performance to Compete with Soft Keyboard

The key issue is whether dictation software for a PDA that processes speech slower than desktop software could be competitive with soft keyboard input. To answer this question, we modeled throughput with processing speeds ranging from real time to 4X. Tables 1 and 2 show how high recognition accuracy must be to be competitive with the soft keyboard, given various processing speeds, assuming correction speeds of 15 and 20 seconds per correction for each speaking rate.

Table 1
Expected throughput rates for the 150-WPM speaker.

Processing Speed	Sec per correction	Recognition Accuracy	Throughput (CWPM)
real time	15	70%	12.24
1.5X	15	75%	14.01
2X	15	75%	13.58
2.5X	15	75%	13.19
3X	15	75%	12.80
3.5X	15	75%	12.46
4X	15	75%	12.12
real time	20	80%	13.64
1.5X	20	80%	13.24
2X	20	80%	12.86
2.5X	20	80%	12.50
3X	20	80%	12.15
3.5X	20	85%	14.75
4X	20	85%	14.29

Table 2
Expected throughput rates for the 100-WPM speaker.

Processing Speed	Sec per correction	Recognition Accuracy	Throughput (CWPM)
real time	15	75%	13.79
1.5X	15	75%	13.19
2X	15	75%	12.63
2.5X	15	75%	12.12
3X	15	80%	13.64
3.5X	15	80%	13.04
4X	15	80%	12.50
real time	20	80%	13.04
1.5X	20	80%	12.50
2X	20	80%	12.00
2.5X	20	85%	14.29
3X	20	85%	13.64
3.5X	20	85%	13.04
4X	20	85%	12.50

The tables show that, even if a dictation input method processes speech at a speed as slow as 4X, this new PDA input method is competitive with soft-keyboard input (the fastest current method).

Considering the Upper Limits of Recognition Accuracy

A review of journalists' estimates of desktop dictation software from 1997 to 1999 suggests that accuracy levels for commercial desktop products are typically around 95% (Lewis, 2001). Due to certain limitations of PDAs (e.g., lack of memory and processing speed; microphone quality) we expect that recognition accuracy would be lower with a PDA. We will assume the upper limit of PDA dictation recognition accuracy to be 90%. Tables 3 and 4 show the expected throughput rates for a 150-WPM speaker and for 100-WPM speaker if a development team could obtain the very high recognition accuracy of 90%. The table displays the throughput rates for correction speeds of 15 and 20 seconds.

Table 3
Expected throughput rates given 90% recognition accuracy for the 150-WPM Speaker

Processing Speed	**Sec per correction**	**Throughput (CWPM)**
real time	15	31.58
1.5X	15	29.51
2X	15	27.69
2.5X	15	26.09
3X	15	24.62
3.5X	15	23.38
4X	15	22.22
real time	20	25.00
1.5X	20	23.68
2X	20	22.50
2.5X	20	21.43
3X	20	20.43
3.5X	20	19.57
4X	20	18.75

This table reveals some very interesting model predictions. It shows that, if recognition accuracies of 90% could be obtained with a PDA dictation system, the dictation input method would be far superior to soft keyboard input, even when running at 4X. If a development team could achieve real time speech processing, users could enter accurate data more than 2 times as quickly as with methods currently available.

This table also demonstrates that correction speed is a much more important determinate of true throughput than speech processing speed. Across correction speeds, as processing time is increased by 50% from real time to 1.5X, throughput is reduced by only 5.99% and as processing speed is increased by an additional 33.33% (from 1.5X to 2X), throughput is reduced by only an additional 5.64%. On the other hand, across processing speeds, as time per correction increases 33.33% (from 15 to 20 sec/correction), throughput decreases by 18.22%.

Table 4
Expected throughput rates given 90% recognition accuracy for the 100-WPM Speaker

Processing Speed	**Sec per correction**	**Throughput (CWPM)**
real time	15	28.57
1.5X	15	26.09
2X	15	24.00
2.5X	15	22.22
3X	15	20.69
3.5X	15	19.35
4X	15	18.18
real time	20	23.08
1.5X	20	21.43
2X	20	20.00
2.5X	20	18.75
3X	20	17.65
3.5X	20	16.67
4X	20	15.79

The data presented in Table 4 mirror that which Table 3 displays, but at a slightly slower rate. As illustrated, even relatively slow speakers using a system with an extremely slow speech processor would benefit greatly from a PDA dictation input method that could offer such a high accuracy level.

General Discussion

Commarford and Lewis (2002) previously showed that dictation could possibly be a very valuable PDA input method, given real-time speech processing. The purpose of creating the current models was to consider whether this value would still be present with more realistic system performance and to provide to practitioners data that suggest how high recognition accuracy must be for a system with a given processing speed to compete with soft keyboard input. The models demonstrate the following key findings:

1. If a recognition accuracy of 85% or greater can be achieved, users will likely be more proficient with a dictation input method than with soft keyboards, even if the system runs at 4X.

2. If 2.5X can be achieved, recognition accuracy need only be 75% for dictation to compete with soft keyboard input.

3. If recognition accuracy is extremely high (90%), a fast speaker will be able to achieve throughput rates that are 1.5-2 times as fast as those obtained with a soft keyboard.

It is possible, however, that an absolute threshold exists for which users find slower speech processing to be unacceptable and frustrating. Perhaps users would not tolerate the need to wait for speech to appear on the screen or the pauses necessary for the system to "catch up." This, however, would not necessarily make such a system unwanted. We must consider that today's users commonly express frustration in response to the tedious nature of currently available PDA input methods.

Final Considerations

The definition we used for nX was for continuous dictation, which is representative of the way users read, but not the way users speak when composing a document or message. When composing, users are likely to pause much more often as they consider what to say next. During these periods of silence, the system would process speech more quickly (2/3nX).

Also, if individuals use a "correct as you go" strategy (rather than dictating a large amount of text before correcting), then the system delay would probably be less noticeable and/or bothersome to users, and this would probably also increase throughput speed. For example, at 150 WPM, a user can speak 20 words in 8 seconds. At 3X it would take 24 seconds before the last of the text appears. However, even with an accuracy rate of 90%, we would expect 2 errors, which would likely take approximately 40 seconds to correct. Therefore as the user is correcting the first error, the system processes the last of the text. In this case it would take 48 seconds to input 20 corrected words. This rate of 25 CWPM can be compared to the expected 20.43 CWPM if a user sped ahead of the system and waited until the end to correct all the misrecognitions. What is possibly more important is the fact that users, who correct as they go, would never experience extremely long lags and would be occupied during times of lag. In fact, this may the natural behavior when the system is unable to keep up. Note, however, that manipulating the GUI will probably slow the processing speed, which would reduce the estimated 25 CWPM throughput by an unknown amount.

References

Commarford, P.M. & Lewis, J. R. (2002). *Models of throughput rates for dictation and voice spelling for handheld computers.* (Tech. Report 29.3544). Raleigh, NC: International Business Machines Corp**.**

Lewis, J. R. (2001). *The accuracy wars: Journalists' estimates of continuous speech product dictation accuracy from 1997-1997.* (Tech. Report 29.3465). Raleigh, NC: International Business Machines Corp**.**

Lewis, J. R. (1999). Effect of error correction strategy on speech dictation throughput. In *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting* (pp. 457-461). Santa Monica, CA: Human Factors Society.

Sears, A., & Arora, R. (2001). An evaluation of gesture recognition for PDAs. In M. J. Smith, G. Salvendy, D. Harris, & R. J. Koubek (Eds.), *Usability Evaluation and Interface Design: Cognitive Engineering, Intelligent Agents and Virtual Reality* (pp. 1-5). Mahwah, NJ: Lawrence Erlbaum Associates.

Zha, Y., & Sears, A. (2001). Data entry for mobile devices using soft keyboards: Understanding the effect of keyboard size. In M. J. Smith, G. Salvendy, D. Harris, & R. J. Koubek (Eds.), *Usability Evaluation and Interface Design: Cognitive Engineering, Intelligent Agents and Virtual Reality* (pp. 16-20). Malwah, NJ: Lawrence Erlbaum Associates.