



## Models of Throughput Rates for Dictation and Voice Spelling for Handheld Devices

PATRICK M. COMMARFORD

*IBM Corporation, 8051 Congress Ave, Suite 2228, Boca Raton, FL 33487, USA*  
commarfo@us.ibm.com

JAMES R. LEWIS

*IBM Corporation, 8051 Congress Ave, Suite 2227, Boca Raton, FL 33487, USA*

**Abstract.** Since the emergence of the personal digital assistant (PDA), developers have attempted to create input methods that allow users to enter accurate data at speeds that approach those achieved with the personal computer. Common text entry methods (handwriting and soft keyboard) allow for rates that are unacceptably slow for many purposes. The objective of this paper is to consider the possible benefits of speech-to-text input mechanisms (dictation and voice spelling) for handheld devices. By modeling throughput based on varying rates of speech, correction speeds, and system recognition accuracies, we can compare expected speech throughput rates to current throughput rates for PDAs.

**Keywords:** personal digital assistant (PDA), dictation, voice spelling, speech interface

### Introduction

Recent years have seen the emergence and rising popularity of handheld personal digital assistants (PDAs). These devices have many benefits, including being small, lightweight, and extremely mobile. To date, there have been two primary methods of inputting data to a PDA—tapping a small onscreen (soft) keyboard and using highly constrained handwriting recognizers such as Graffiti (Graffiti is a registered trademark of Palm Computing, Inc.) or Unistrokes (Unistrokes is a registered trademark of Xerox Corp.). The current input speeds for these methods, however, are substantially slower than input rates achieved with a personal computer and keyboard.

Virtually all users of PDAs have experience with personal computers and have some familiarity with the standard computer keyboard (the QWERTY keyboard), with which expert typists can enter data at a rates of approximately 55 words per minute (WPM) with near perfect accuracy (Marklin et al., 1998; Norman and Fisher, 1982). Prior research (discussed

below) has shown substantially slower rates of input for various handwriting recognizers and soft keyboards.

Hand printing speeds are typically in the 12–23 WPM range, and cursive handwriting speeds range from 16 to just over 30 WPM (Soukoreff and MacKenzie, 1995). These, necessarily, provide estimates of the upper limits for this sort of text entry. MacKenzie and Chang (1999), using two discrete printing recognizers and a 9.5-inch tablet, found a mean text entry speed of 17.1 WPM. The mean recognition accuracy was 92% when the recognizer was constrained to lowercase letters and 90% when constrained to upper and lower case letters. MacKenzie and Zhang (1997), found high (95.8%) recognition accuracy for experienced users of the Graffiti handwriting recognition system, but did not report the entry speeds. Sears and Arora (2001) reported a much slower text entry rate of 4.95 WPM for the Graffiti recognizer with a recognition accuracy of 95% for participants using a PDA (rather than a tablet). They found that participants using the Jot recognizer were able to produce 7.74 WPM with an average recognition accuracy of 88% (Jot is

a registered trademark of Communication Intelligence Corporation). Each of the previously mentioned studies reported rates of entry for uncorrected text. Kleid and Bonto (1995) asked users to enter a rather complex set of letters, numbers, and special characters (a person's contact information) using Graffiti on a 6" (diagonal) screen. Participants were to attempt 100% accuracy and to use any editing tools that they felt would be helpful. The participants were only able to enter 1.98 corrected words per minute (CWPM).

Research has shown stylus tapping on a soft QWERTY keyboard to be slightly faster than using a handwriting recognizer, but these rates are still sub-optimal. Zha and Sears (2001) reported that participants could input text on a PDA at a rate of 12.62 WPM with an average input accuracy of 96% using the soft keyboard. Using a tablet, MacKenzie et al. (1994) found that users could type text at a rate of 22.9 WPM with 99% accuracy. MacKenzie et al. (1999) reported text entry rates of 20.2 WPM for participants tapping on a full-sized paper QWERTY layout. Kleid and Bonto (1995) had their participants use a soft keyboard to enter the previously described set of complex text with 100% accuracy. Under these conditions, participants were only able to obtain mean throughput rates of 5.17 CWPM.

In the past decade, users have been introduced to a new method of inputting data to a personal computer—speech dictation. While voice throughput is not yet typically as fast as typing with a full-sized QWERTY keyboard, it may offer a faster rate of throughput for handheld devices than the currently available options. Lewis (1999) defined true throughput as the number of correct words produced per minute, and found that participants could achieve rates of 31.0 CWPM with multimodal (manual and vocal) correction and 19.0 CWPM with voice-only correction, using two commercially available desktop speech dictation products. Voice-only correction was significantly slower (29.1 seconds per correction) than multimodal correction (13.2 seconds per correction).

If system designers could embed voice recognition technology into the PDA environment in a way that allows users to achieve throughput rates similar to those observed by Lewis (1999), users would benefit greatly. By 1999, estimates of recognition accuracy for commercial desktop dictation systems ranged from 90–95% (Koester, 2001; Lewis, 2001). Resource limitations of these handheld devices, however, will probably prevent the high levels of recognition accuracy reached

by the desktop software. In addition, multimodal correction on a PDA will include the use of a method of input (soft keyboard or handwriting recognizer) that is less efficient than those used with the desktop systems in the Lewis study.

It is reasonable to consider the efficiency of two methods of inputting data by voice in a PDA environment: speech dictation and voice spelling. These methods can exist in isolation or in combination. To dictate, users would simply say the words they want to appear on the screen. To voice spell, the user would say codes for each letter of the alphabet. Due to the well-known acoustic similarities of certain letters (for example, "B" and "D" or "S" and "F"), recognition accuracy for letter names is much lower than for a carefully selected set of code words (Lewis and Commarford, 2002). For example, rather than saying "d" the user would say, "dog." It would be possible to present a reminder display when in voice spelling mode for users who do not have the codes committed to memory.

The user would certainly be able to produce more uncorrected words per minute via dictation than by voice spelling. However, because of the limited grammar set, voice spelling would probably achieve higher levels of system recognition accuracy than dictation, potentially reducing the need to correct. Furthermore, corrections speeds for voice spelling (single-character corrections) would probably be faster than correction speeds for dictation (full-word corrections). Prior modeling (Lewis, 1999) suggests that system recognition accuracy and correction speeds are more important determinants of true throughput than speaking rate. Modeling true throughput (CWPM) based on input method, system accuracy, time per correction, and speaking rate can help us understand how substantial the accuracy difference between spelling and dictation would have to be for it to be advantageous to use the spell mode. These models would also allow comparisons of each of these speech input methods to reported throughput rates for other PDA input methods.

The aforementioned research suggests faster rates of throughput for soft keyboard input than for handwriting recognizers with a PDA. The Zha and Sears (2001) study seems to provide the best estimate of PDA throughput with a soft keyboard. These researchers used a PDA and asked participants to input a 41-word passage characteristic of a short business email message. They found a mean input rate for this task of 12.62 WPM, with a 4% error rate. Assuming a very high speed of correction, it seems reasonable to set the

benchmark for the true throughput rate for soft keyboard input at 12 CWPM.

The rest of this report describes performance modeling conducted to make the following comparisons of throughput rates:

- (a) dictation for the 150-WPM speaker vs. the 100-WPM speaker
- (b) expert spelling for the 150-WPM speaker vs. the 100-WPM speaker
- (c) novice spelling for the 150-WPM speaker vs. the 100-WPM speaker
- (d) dictation vs. expert spelling for the 150-WPM speaker
- (e) dictation vs. expert spelling for the 100-WPM speaker

In addition, there will be a comparison of each speech input rate to the target rate of 12 CWPM, which appears to be the best estimate of the most efficient method of text input currently available for stand-alone PDAs (not docked and connected to a personal computer).

## Method

The source text for this evaluation was a 107-word passage with 483 letters. This passage contained 101 spaces and seven punctuation marks. The passage contained no capitalization, except for characters following periods. We created six models of true throughput (using CWPM) for speech dictation and voice spelling (see Table 1). Each model contains a range of system recognition accuracies from 50 to 100% and a range of correction times from 5 to 35 seconds.

The expert voice spelling models assume complete automaticity in the assignment of the letters to their

respective codes (in other words, users need no processing time to match letters to their codes or to retrieve them from short term memory). The novice models assume a 230 ms eye movement time to locate each letter in the passage on the spell vocabulary reminder display. The 230 ms eye movement time is consistent with that given as the typical or “middle man” time by Card et al. (1983).

## The Models

### *Dictation Model*

This model is a replication of that created by Lewis (1999), but extends the upper limit for correction times to 35 s because it is reasonable to anticipate longer correction times with a handheld device. Table 2 shows the expected throughput data (in CWPM) for a 150-WPM speaker and a 100-WPM speaker at varying levels of system recognition accuracy and varying correction speeds. The bold numbers in the table indicate the point at which speech becomes competitive with soft keyboard input for each combination of speaking rate, correction speed, and recognition accuracy (in other words, exceed 12 CWPM). Figure 1 illustrates this relationship.

The following worked example illustrates the method by which we calculated CWPM for dictation. The example is based on the cell in Table 2 that provides the throughput rate for a 150-WPM speaker who takes 25 seconds on average to complete a correction and is using a system that accurately recognizes 80% of the dictated speech.

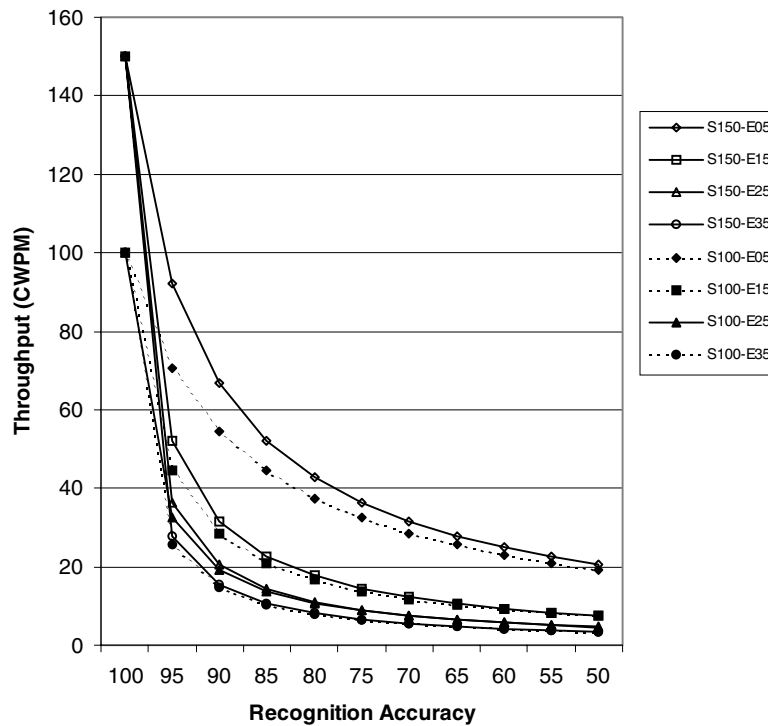
We calculated CWPM by dividing the number of words to appear in the target application on the PDA screen (107 in this case) by the total amount of time it would take, in minutes, for the corrected text to appear. Total time consists of time to speak the passage and time to correct the misrecognitions. We calculated time to speak by dividing the number of words to be spoken (107) by the speaking rate (150 WPM). The result was 0.71 minutes to speak the 107-word passage. Multiplying the number of required corrections by the correction speed (25 seconds) provided the correction time. We determined the number of required corrections by multiplying the number of words spoken (107) by the error rate ( $1 - \text{recognition accuracy}$ ). For the example data, the error rate is .20, resulting in 21.4 errors. Multiplying 21.4 errors by the 25-second correction speed produced 535 seconds, or 8.92 minutes. The

Table 1. Description of six throughput models.

Input method	User	Speaking rate
Dictation	All	100 WPM
		150 WPM
Voice spelling	Novice	100 WPM
		150 WPM
	Expert	100 WPM
		150 WPM

Table 2. Model of expected throughput (CWPM) rates for dictation.

Speaking rate	Correction speed	Recognition accuracy										
		100%	95%	90%	85%	80%	75%	70%	65%	60%	55%	50%
150 WPM	5 sec	150	92.31	66.67	52.17	42.86	36.36	31.58	27.91	25.00	22.64	<b>20.69</b>
	15 sec	150	52.17	31.58	22.64	17.65	14.46	<b>12.24</b>	10.62	9.38	8.39	7.59
	25 sec	150	36.36	20.69	<b>14.46</b>	11.11	9.02	7.59	6.56	5.77	5.15	4.65
	35 sec	150	27.91	<b>15.38</b>	10.62	8.11	6.56	5.50	4.74	4.17	3.72	3.35
100 WPM	5 sec	100	70.59	54.55	44.44	37.50	32.43	28.57	25.53	23.08	21.05	<b>19.35</b>
	15 sec	100	44.44	28.57	21.05	16.67	<b>13.79</b>	11.76	10.26	9.09	8.16	7.41
	25 sec	100	32.43	19.35	<b>13.79</b>	10.71	8.76	7.41	6.42	5.66	5.06	4.58
	35 sec	100	25.53	<b>14.63</b>	10.26	7.89	6.42	5.41	4.67	4.11	3.67	3.31



S150 = 150 WPM speaking rate  
 S100 = 100 WPM speaking rate  
 E05 = 5 seconds per correction  
 E15 = 15 seconds per correction  
 E25 = 25 seconds per correction  
 E35 = 35 seconds per correction

Figure 1. Model of expected throughput (CWPM) rates for dictation.

sum of 0.71 minutes (speaking time) and 8.92 minutes (correction time) was 9.63 minutes (total time). Finally, dividing the 107 corrected words by 9.63 minutes of total time produced the presented value of 11.11 CWPM (see Table 2).

Referring to Table 2, beginning with the right hand side of the table, we can see that with a recognition accuracy of 50%, the user would need to be able to make a correction every five seconds for dictation to compete with soft keyboard input. This speed

of correction seems unrealistically fast given Lewis' (1999) research, in which he found average multimodal correction speeds of 13.2 seconds per correction with desktop speech dictation systems. Given the size and processing limitations of the PDA, correction times for dictation would likely be a little slower, perhaps between 15 and 20 seconds. If we assume 15 seconds per correction, recognition accuracies as low as 70–75% would produce mean true throughput rates that are highly competitive with the target of 12 CWPM, even for a 100-WPM speaker. If correction speeds were as slow as 25 seconds per correction, a recognition accuracy of about 85% would be necessary for dictation to compete with soft keyboard input. At 35 seconds per correction, the target accuracy would be 90%.

### Expert Speller Model

The expert speller model allows estimation of the voice spelling throughput rates for a speaker for whom the spell letter codes have become automatic. Table 3 shows the expected spell data for an expert speller, speaking 150 and 100 WPM at varying levels of recognition accuracy and varying correction speeds. The bold numbers in the table indicate the point at which speech becomes competitive with soft keyboard input. Figure 2 illustrates this relationship.

The calculations for the expert spelling model are similar to those for the dictation model, except that the number of words spoken is now 591 (the user must say a code word for each of 483 letters, must say, "space" 101 times, and must say, "period" 7 times) and corrections are at the character rather than the word level. The following example goes through the steps of calculating throughput for an expert speller speaking 150-WPM, correcting one character every 25 seconds, and using a

system that accurately recognizes 80% of the speech.

Dividing the number of utterances to be spoken (591) by the 150-WPM speaking rate produced 3.94 minutes to speak the passage. Multiplying the number of utterances spoken (591) by the error rate (.20) produced 118.2 errors, which would require 2955 seconds (49.25 minutes) to correct at 25 seconds per error. Summing the time to speak (3.94 minutes) and the time to correct (49.25 minutes) produced a total time of 53.19 minutes. Dividing 107 words in the passage by 53.19 produced 2.01 CWPM (see Table 3).

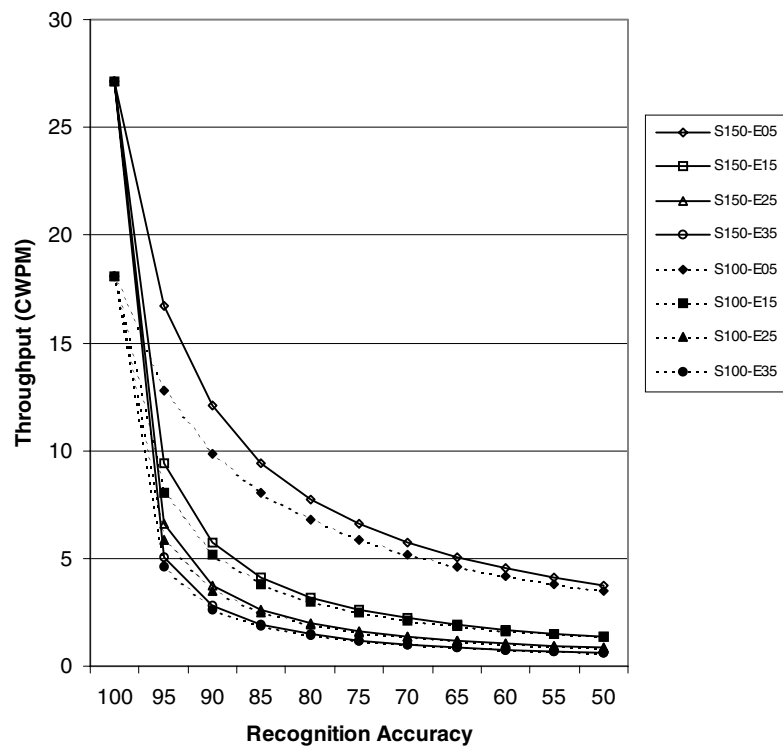
Table 3 shows that, with low levels of system recognition accuracy, voice spelling will produce unacceptably slow input rates, regardless of correction speed. Assuming a correction speed of 15 seconds, recognition accuracy would need to be greater than 95% for voice spelling to be a competitive method of inputting text. However, the previously cited correction speeds (Lewis, 1999) were for full-word corrections that resulted from system misrecognitions of user dictation. Correcting the misrecognition of a single character would presumably be much quicker (see Koester, 2001, p. 123), requiring only two actions (tapping or saying backspace, followed by tapping the appropriate key or saying the appropriate code). Assuming five seconds per correction, 90% recognition accuracy would allow 150-WPM speakers to voice spell at a rate competitive with soft keyboard input (12.07 CWPM). Slower speaking expert spellers would require a recognition accuracy of 95% for voice spelling to be a competitive alternative.

### Novice Speller Model

The novice speller model allows the estimation of the throughput rates for a speaker who is just learning to

Table 3. Model of expected throughput (CWPM) rates for the expert speller.

Speaking rate	Correction speed	Recognition accuracy										
		100%	95%	90%	85%	80%	75%	70%	65%	60%	55%	50%
150 WPM	5 sec	27.16	16.71	<b>12.07</b>	9.45	7.76	6.58	5.72	5.05	4.53	4.10	3.75
	15 sec	<b>27.16</b>	9.45	5.72	4.10	3.19	2.62	2.22	1.92	1.70	1.52	1.38
	25 sec	<b>27.16</b>	6.58	3.75	2.62	2.01	1.63	1.38	1.19	1.04	0.93	0.84
	35 sec	<b>27.16</b>	5.05	2.79	1.92	1.47	1.19	1.00	0.86	0.75	0.67	0.61
100 WPM	5 sec	18.10	<b>12.78</b>	9.88	8.05	6.79	5.87	5.17	4.62	4.18	3.81	3.50
	15 sec	<b>18.10</b>	8.05	5.17	3.81	3.02	2.50	2.13	1.86	1.65	1.48	1.34
	25 sec	<b>18.10</b>	5.87	3.50	2.50	1.94	1.59	1.34	1.16	1.02	0.92	0.83
	35 sec	<b>18.10</b>	4.62	2.65	1.86	1.43	1.16	0.98	0.85	0.74	0.66	0.60



S150 = 150 WPM speaking rate  
 S100 = 100 WPM speaking rate  
 E05 = 5 seconds per correction  
 E15 = 15 seconds per correction  
 E25 = 25 seconds per correction  
 E35 = 35 seconds per correction

Figure 2. Model of expected throughput (CWPM) rates for the expert speller.

use spell mode and has not yet memorized the letter codes. Table 4 shows the expected throughputs for a novice speller at 150 WPM and at 100 WPM with varying levels of system recognition accuracy and varying correction speeds. The bold numbers in the table indicate the point at which speech becomes competitive with soft keyboard input. Figure 3 illustrates this relationship.

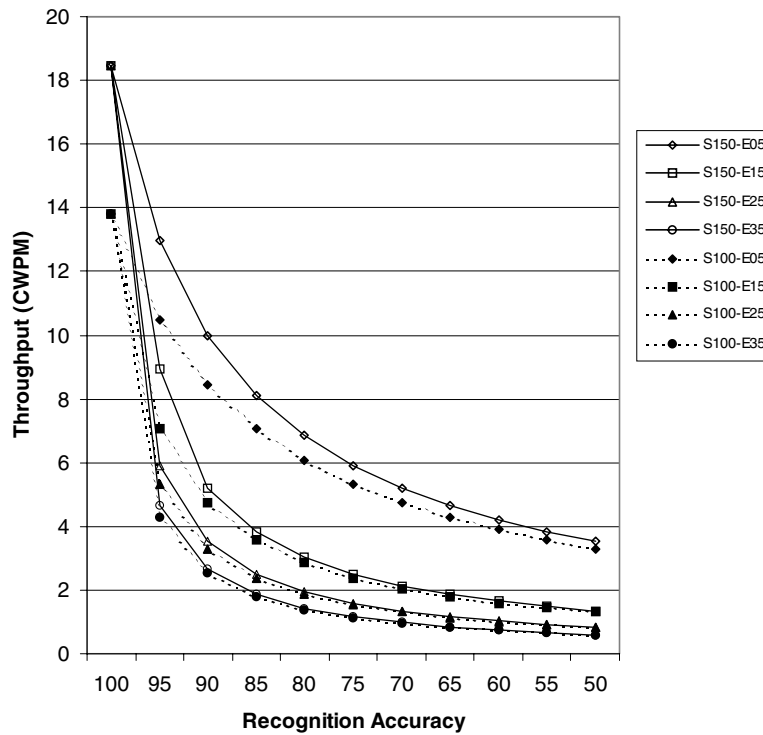
We calculated novice speller throughput times in much the same way as the expert speller models. The one exception is the addition of a time component, code word search time, to the total time. The following describes the method of calculating throughput for a novice speller speaking 150-WPM, requiring 25 seconds per correction for each misrecognized character, and using a system that accurately recognizes 80% of the spoken words.

The 3.94 minutes to speak the passage and the 49.25 minutes to correct the misrecognitions remain the same as in the expert speller example above. However, we also assumed a 230 msec (.0038 minute) search time for code words for each of the 483 letters in the passage (but not for "period" or "space"). Multiplying 483 code words by 0.0038 minutes per search yielded 1.85 minutes of total search time. Summing time to speak (3.94 minutes), time to correct (49.25 minutes), and time to search (1.85 minutes) resulted in 55.04 minutes. Dividing 107 by 55.04 minutes produced 1.94 CWPM (see Table 4).

The expected novice speller data follows a pattern similar to the expert speller, but is slightly slower. At five seconds per correction, novice spellers would require 95% or greater recognition accuracy to be competitive with the soft keyboard.

Table 4. Model of expected throughput (CWPM) rates for the novice speller.

Speaking rate	Correction speed	Recognition accuracy										
		100%	95%	90%	85%	80%	75%	70%	65%	60%	55%	50%
150 WPM	5 sec	18.47	<b>12.96</b>	9.98	8.12	6.84	5.91	5.20	4.65	4.20	3.83	3.52
	15 sec	<b>18.47</b>	8.92	5.20	3.83	3.03	2.50	2.14	1.86	1.65	1.48	1.34
	25 sec	<b>18.47</b>	5.91	3.52	2.50	1.94	1.59	1.34	1.16	1.03	0.92	0.83
	35 sec	<b>18.47</b>	4.65	2.66	1.86	1.43	1.16	0.98	0.85	0.74	0.66	0.60
100 WPM	5 sec	<b>13.79</b>	10.47	8.43	7.06	6.08	5.33	4.75	4.28	3.90	3.58	3.30
	15 sec	<b>13.79</b>	7.06	4.75	3.58	2.87	2.39	2.05	1.80	1.60	1.44	1.31
	25 sec	<b>13.79</b>	5.33	3.30	2.39	1.88	1.54	1.31	1.14	1.01	0.90	0.82
	35 sec	<b>13.79</b>	4.28	2.53	1.80	1.39	1.14	0.96	0.83	0.73	0.66	0.59



S150 = 150 WPM speaking rate  
 S100 = 100 WPM speaking rate  
 E05 = 5 seconds per correction  
 E15 = 15 seconds per correction  
 E25 = 25 seconds per correction  
 E35 = 35 seconds per correction

Figure 3. Model of expected throughput (CWPM) rates for the novice speller.

150-WPM Speaker Model

This model provides expected throughputs for someone who speaks at a rate of 150 WPM for dictation and for expert spelling. The data enable the determi-

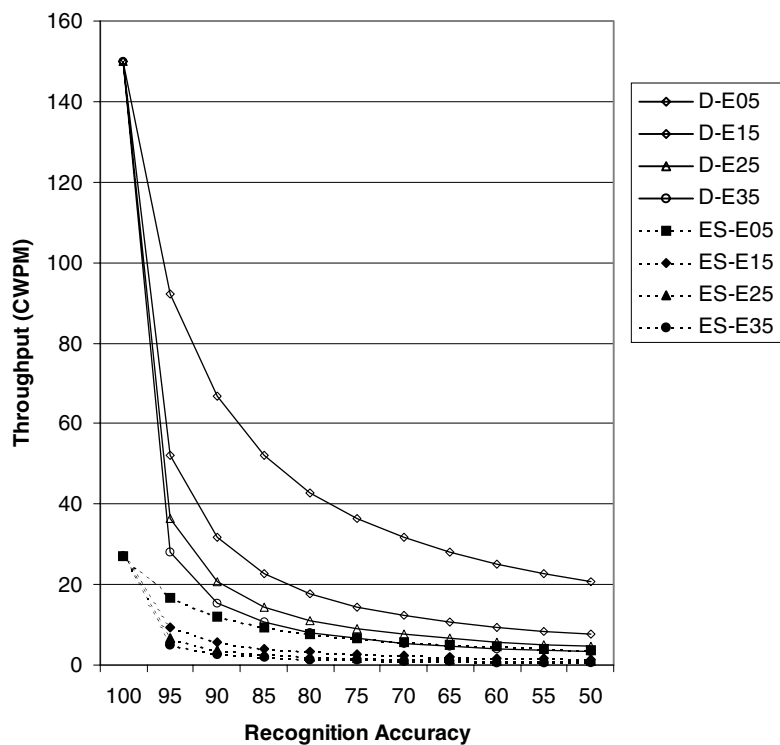
nation, for a given correction speed, of how much difference in recognition accuracy must exist for expert voice spelling to be more efficient than dictation. Table 5 shows the expected dictation and expert spelling throughput rates for a user who speaks at a rate of

Table 5. Model of expected throughput (CWPM) rates for the 150-WPM speaker.

Input method	Correction speed	Recognition accuracy										
		100%	95%	90%	85%	80%	75%	70%	65%	60%	55%	50%
Dictation	5 sec	150.00	92.31	66.67	52.17	42.86	36.36	31.58	27.91	25.00	22.64	<b>20.69</b>
	15 sec	150.00	52.17	31.58	22.64	17.65	14.46	<b>12.24</b>	10.62	9.38	8.39	7.59
	25 sec	150.00	36.36	20.69	<b>14.46</b>	11.11	9.02	7.59	6.56	5.77	5.15	4.65
	35 sec	150.00	27.91	<b>15.38</b>	10.62	8.11	6.56	5.50	4.74	4.17	3.72	3.35
Spelling	5 sec	27.16	16.71	<b>12.07</b>	9.45	7.76	6.58	5.72	5.05	4.53	4.10	3.75
	15 sec	<b>27.16</b>	9.45	5.72	4.10	3.19	2.62	2.22	1.92	1.70	1.52	1.38
	25 sec	<b>27.16</b>	6.58	3.75	2.62	2.01	1.63	1.38	1.19	1.04	0.93	0.84
	35 sec	<b>27.16</b>	5.05	2.79	1.92	1.47	1.19	1.00	0.86	0.75	0.67	0.61

150 WPM for varying levels of system recognition accuracy and varying correction speeds. The bold numbers in the table indicate the point at which speech becomes competitive with soft keyboard input. Figure 4 illustrates this relationship.

As shown in Table 5, given equivalent system recognition accuracy, dictation will be superior across all correction speeds, with the greatest differences coming at the higher levels of accuracy. Given the large difference in the size of the grammar sets for the two



D = Dictation  
 ES = Expert Spelling  
 E05 = 5 seconds per correction  
 E15 = 15 seconds per correction  
 E25 = 25 seconds per correction  
 E35 = 35 seconds per correction

Figure 4. Model of expected throughput (CWPM) rates for the 150 WPM speaker.

modes, however, recognition accuracy is expected to be higher and correction speeds faster when the user is in spell mode. Lewis and Commarford (2002) developed a voice spelling alphabet and tested its accuracy with a desktop system and headset microphone. The grammar set (which included letter codes for voice spelling, punctuation, and cursor control commands) produced results that were 97.5% accurate under these conditions. Given that accuracy should be no higher (and might be lower) with a PDA microphone, 97.5% is an estimate of the upper limit to voice spelling accuracy. Assuming 95% recognition accuracy and 5 seconds per correction for voice spelling and assuming 15 seconds per correction for dictation, dictation with recognition accuracy as low as 80% would be more efficient than voice spelling (with a voice spelling throughput of 16.71 CWPM and a dictation throughput of 17.65 CWPM).

#### 100-WPM Speaker Model

This model provides expected dictation and voice spelling throughput rates for an expert speller who speaks at a rate of 100 WPM. Table 6 shows the expected throughput rates. The bold numbers in the table indicate the point at which speech becomes competitive with soft keyboard input. Figure 5 illustrates the relationship.

Again, the data show that, given equivalent system recognition accuracy, dictation should be superior to voice spelling across all correction speeds, with the greatest differences at the higher levels of accuracy. Assuming 95% recognition accuracy and 5 seconds per correction for voice spelling and assuming 15 seconds per correction for dictation, dictation with recognition accuracy as low as 75% would be more efficient

than voice spelling (with a voice spelling throughput of 12.78 CWPM and a dictation throughput of 13.79 CWPM).

### General Discussion

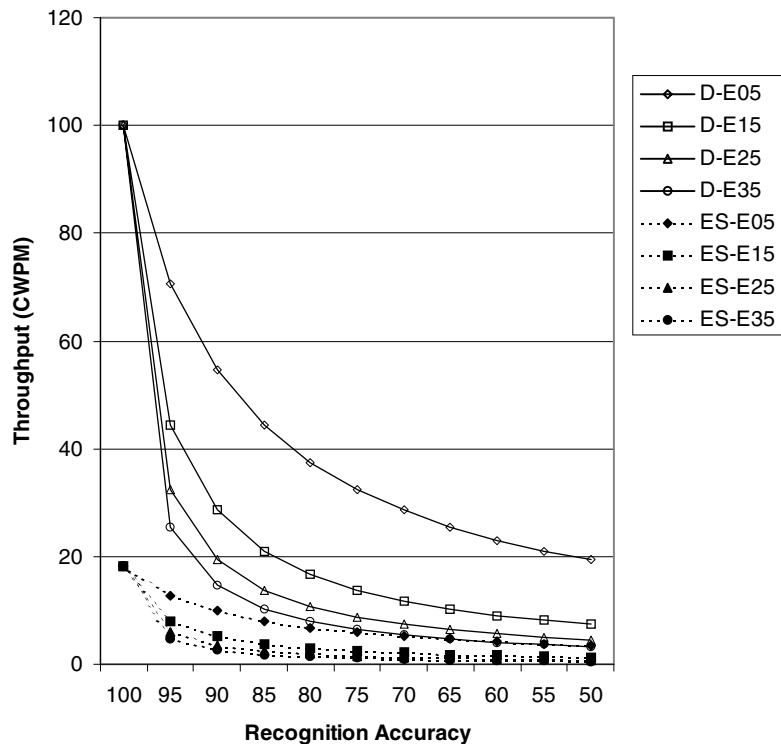
#### *Relatively Low Throughput for Handwriting Recognition*

Although it might seem counterintuitive, the evidence from the published literature indicates that textual throughput with a soft keyboard is more efficient than handwriting recognition for PDAs, despite the improvements in recognition accuracy associated with constrained alphabets such as Graffiti and Unistrokes. The only cause of error in key tapping is tapping the wrong key, but errors in handwriting recognition can be due to forming an incorrect character (user error) or misrecognition of a correctly formed character (system error). Furthermore, corrections performed with handwriting recognition are also probabilistic, making them more error prone than deterministic key tapping. Thus, the low throughput observed for handwriting recognition systems is most likely the result of a longer total correction time. Regardless of the cause, the throughputs reported in the literature indicate that handwriting recognition is a less effective input method than typing with soft keyboards.

For these reasons, we have not included handwriting throughput in our models. Instead, we have focused on comparisons of speech input methods with the more competitive method of soft keyboard input. For readers interested in comparing speech throughput rates with handwriting on a PDA, a generous estimate for the true throughput of current handwriting recognition

Table 6. Model of expected throughput (CWPM) rates for the 100-WPM speaker.

Input method	Correction speed	Recognition accuracy										
		100%	95%	90%	85%	80%	75%	70%	65%	60%	55%	50%
Dictation	5 sec	100.00	70.59	54.55	44.44	37.50	32.43	28.57	25.53	23.08	21.05	<b>19.35</b>
	15 sec	100.00	44.44	28.57	21.05	16.67	<b>13.79</b>	11.76	10.26	9.09	8.16	7.41
	25 sec	100.00	32.43	19.35	<b>13.79</b>	10.71	8.76	7.41	6.42	5.66	5.06	4.58
	35 sec	100.00	25.53	<b>14.63</b>	10.26	7.89	6.42	5.41	4.67	4.11	3.67	3.31
Spelling	5 sec	18.10	<b>12.78</b>	9.88	8.05	6.79	5.87	5.17	4.62	4.18	3.81	3.50
	15 sec	<b>18.10</b>	8.05	5.17	3.81	3.02	2.50	2.13	1.86	1.65	1.48	1.34
	25 sec	<b>18.10</b>	5.87	3.50	2.50	1.94	1.59	1.34	1.16	1.02	0.92	0.83
	35 sec	<b>18.10</b>	4.62	2.65	1.86	1.43	1.16	0.98	0.85	0.74	0.66	0.60



D = Dictation  
 ES = Expert Spelling  
 E05 = 5 seconds per correction  
 E15 = 15 seconds per correction  
 E25 = 25 seconds per correction  
 E35 = 35 seconds per correction

Figure 5. Model of expected throughput (CWPM) rates for the 100 WPM speaker.

is 7 CWPM (based on the results for the Jot recognizer in Sears and Arora, 2001).

#### *Potential Usefulness of Voice Spelling and Dictation as PDA Input Methods*

The range of values used to model true throughput for the speech input methods described in this paper have a firm basis in previously published human performance data, and were selected to encompass them. The typical speaking rate for users who are transcribing text into a dictation system is about 110 WPM (Lewis, 1999), so we set the values for this rate in the models at 100 and 150 WPM.

Correction speeds ranged from about 13 to 30 seconds per correction in Lewis (1999), depending on the correction strategy employed by users (multimodal for the faster times, voice only for the slower times). Correction speeds for entry with standard key-

boards are faster, averaging about 3 seconds per correction (Koester, 2001). Given a soft rather than standard keyboard, it therefore seems reasonable to assume about 5 seconds per correction for individual characters, such as those produced during voice spelling. Thus, the range of correction speeds included in our models (5 to 35 seconds per correction) encompasses correction speeds for keyed corrections, multimodal corrections, and voice-only corrections.

Estimates of recognition accuracy for commercial desktop speech dictation systems typically range from 90 to 95%, but with some additional variation above and below that range (Koester, 2001; Lewis, 2001). Resource and hardware limitations might prevent users from achieving such high recognition accuracies with a PDA. Therefore, the range of recognition accuracies modeled (50 to 100%) seems appropriate to consider.

The models clearly show that both voice spelling and dictation can be competitive with soft keyboard entry,

and are therefore potentially useful methods for the input of text into PDAs. All other things being equal, dictation will be much more effective than voice spelling. Due to their differing system requirements, however, voice spelling might be the only viable method for speech input of text into handheld devices with relatively low system resources.

The models also illustrate the likely limits to dictation throughput as a function of accuracy and correction speed. Users can easily speak text at rates between 100 and 150 WPM, but are very unlikely to experience true throughput much greater than 45 to 50 CWPM (given 95% accuracy and multimodal correction speeds of 15 seconds per correction) in the near future.

#### Summary of Key Findings

1. Assuming 15 seconds per correction with a PDA, dictation would be as efficient as soft keyboard input as long as speech recognition accuracy were 70% or greater. Assuming dictation recognition accuracies as high as 85–90%, dictation would be approximately twice as productive as soft keyboard entry.
2. Under ideal conditions (expert speller, 150-WPM speaker, and 5-second correction speed), voice-spelling accuracy must exceed 90% to be competitive with soft keyboard entry.
3. Voice spelling will not be as efficient as dictation unless voice spelling recognition accuracy is much higher than dictation recognition accuracy.

#### Conclusions

Current methods for textual input to a PDA do not result in a very high throughput. Of the two most common contemporary methods, typing on a virtual keyboard has a higher throughput than any approach to handwriting recognition. The best current estimate of true throughput for typing on a PDA's virtual keyboard is about 12 CWPM. The models presented in this paper show that dictation on a PDA would have a substantial throughput advantage over current methods. Voice spelling is a viable alternative to current methods for PDAs that do not have sufficient computing power to support dictation. The true throughput for any speech input method with less than perfect recognition accuracy is critically dependent on correction speed, making the design of efficient correction procedures of utmost importance.

#### References

- Card, S.K., Moran, T.P., and Newell, A. (1983). *The Psychology of Human Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum.
- Kleid, N.A. and Bonto, M.A. (1995). *Handwriting Recognition and Soft Keyboard Study*. (Tech. Report 29.3008). Raleigh, NC: International Business Machines Corp.
- Koester, H.H. (2001). User performance with speech recognition: A literature review. *Assistive Technology*, 13:116–130.
- Lewis, J.R. (1999). Effect of error correction strategy on speech dictation throughput. *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting*, Santa Monica, CA: Human Factors Society, pp. 457–461.
- Lewis, J.R. (2001). *The Accuracy Wars: Journalists' Estimates of Continuous Speech Product Dictation Accuracy from 1997–1999* (Tech. Report 29.3465). Raleigh, NC: International Business Machines Corp.
- Lewis, J.R. and Commarford, P.M. (2002). *Developing and Tuning a Voice Spelling Alphabet for Devices with Small Displays*. (Tech. Report 29.3517). Raleigh, NC: International Business Machines Corp.
- MacKenzie, S.I. and Chang, L. (1999). A performance comparison of two handwriting recognizers. *Interacting with Computers*, 11:283–297.
- MacKenzie, S.I., Nonnecke, R.B., McQueen, J.C., Riddersma, S., and Meltz, M. (1994). A comparison of three methods of character entry on pen-based computers. *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting*, Santa Monica, CA: Human Factors Society, pp. 330–334.
- MacKenzie, S.I. and Zhang, S.X. (1997). The immediate usability of Graffiti. *Proceedings of Graphics Interface '97*, Toronto, Canada: Canadian Information Processing Society, pp. 129–137.
- MacKenzie, S.I., Zhang, S.X., and Soukoreff, R.W. (1999). Text entry using soft keyboards. *Behaviour and Information Technology*, 18:235–244.
- Marklin, R.W., Simoneau, G.G., and Hoffman, D. (1998). Effects of computer keyboard setup parameters and user's anthropometric characteristics on wrist deviation and typing efficiency. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, Santa Monica, CA: Human Factors Society, pp. 876–880.
- Norman, D.A. and Fisher, D. (1982). Why alphabetic keyboards are not easy to use: Keyboard layout doesn't much matter. *Human Factors*, 24:509–519.
- Sears, A. and Arora, R. (2001). An evaluation of gesture recognition for PDAs. In M.J. Smith, G. Salvendy, D. Harris, and R.J. Koubek (Eds.), *Usability Evaluation and Interface Design: Cognitive Engineering, Intelligent Agents and Virtual Reality*, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 1–5.
- Soukoreff, R.W. and MacKenzie, S.I. (1995). Theoretical upper and lower bounds on typing speed using a stylus and soft keyboard. *Behaviour & Information Technology*, 14:370–379.
- Zha, Y. and 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, and R.J. Koubek (Eds.), *Usability Evaluation and Interface Design: Cognitive Engineering, Intelligent Agents and Virtual Reality*, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 16–20.