Modelling Data Entry Rates for ASR and Alternative Input Methods

Roger K. Moore
Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, UK
r.k.moore@dcs.shef.ac.uk

Abstract
An often-cited advantage of automatic speech recognition (ASR) is that it is ‘fast’; it is quite easy for a person to speak at several hundred words a minute, well above the rates that are possible using other modes of data entry. However, in order to conduct a fair comparison between alternative data entry methods, it is necessary to consider not the input rate per se, but the rate at which it is possible to enter information that is fully correct. This paper describes a model for predicting the relative success of alternative method of data entry in terms of the effective ‘throughput’ that is achievable taking into account typical input data entry rates, error rates and error correction times. Results are presented for the entry of both conventional and SMS-style text.

1. Introduction
Automatic Speech Recognition (ASR) is often hailed as the most ‘natural’ method by which a human could communicate with a machine. Whilst it can be debated whether talking to a machine could ever be viewed as natural, it is nevertheless true that ASR allows a human user to draw on ‘intuitive’ behaviours that have been refined over many years in the context of spoken human-to-human interaction.

It is also the case that spoken language can provide a valuable extra communication channel between a human and a machine in hands/eyes-busy situations - giving significant benefits in terms of reduced workloads and increased safety. As evidence of this, recent years have seen a substantial growth in mobile applications for embedded ASR systems, and this has in part been fueled by the legislation that is now in place in many countries around the world that bans drivers from using mobile phones in cars.

Another often-cited ‘advantage’ of ASR is that it is fast, i.e. it is quite easy for a person to speak at several hundred words a minute, well above the rates that are possible using other modes of data entry. However, in order to conduct a fair comparison between alternative methods, it is necessary to consider not the input rate per se, but the rate at which it is possible to enter information that is fully correct. In other words, since most data entry methods are not error-free, their relative merits can only be judged by taking into account the time it takes to correct any input errors.

The need to perform this type of comparison was illustrated by a study in which the performance of four commercial off-the-shelf speech dictation systems was measured, not just in terms of word error rate, but also in terms of ‘document creation rate’ [1]. As expected, the experiment revealed an inverse correlation between the two - the lower the error rate, the higher the document creation rate.

However, what the comparison also revealed was that, because of the time it took to correct errors, a two-fingered typist could achieve better data entry rates than a person using ASR (see Figure 1). This result provided a partial explanation as to why such dictation systems proved to be of most benefit to those users who either could not or would not type, i.e. more conventional users found that it was faster to stick to the keyboard.

1. Alternative input methods
A range of speaking, typing, tapping, keying and writing methods of data entry were considered in this study:
- large-vocabulary continuous speech recognition
  (e.g. “The cat sat on the mat”)
- ASR of spelled words using the conventional orthographic alphabet
  (e.g. “t”, “h”, “e”, “space”, “c”, “a”, “r”, … etc.)
- ASR of words spelt using the ICAO phonetic alphabet
  (e.g. “tango”, “hotel”, “echo”, “space”, “charlie”, etc)
- typing on a conventional QWERTY keyboard

Figure 1: A comparison between a keyboard and four speech dictation systems as alternative methods for error-free document creation (taken from [1]).

This paper represents an attempt to formalize these relationships by means of a quantitative model. A range of methods for data entry, including ASR, have been investigated, and results from the model are presented that predict the relative success of each method of data entry in terms of the effective ‘throughput’ that is obtainable. The model has been calibrated on typical data input entry rates, error rates and error correction times.
• soft typing, i.e. tapping the soft-keyboard of a PDA using a stylus
• multi press, i.e. pressing keys on a mobile phone keypad
• using T9® ‘text completion’ on a mobile phone keypad
• handwriting recognition

The speech-based spelling methods were included in order to determine whether the higher accuracy obtainable using such small vocabularies (as compared to a large vocabulary dictation system) could mitigate the fact that spelling is inherently slower than speaking normally. The PDA and mobile phone based modes were included in order to ascertain whether speech-input offered an advantage over such small and fiddly keyboards.

For the speech-based input modalities, the investigation also considered both manual and hands-free (i.e. point-and-click vs. spoken) modes for a user to perform the necessary error corrections.

3. A model for data entry

A model for data entry that takes into account the effects of errors and error correction on the ‘true’ throughput rate has been introduced by Lewis [2]. The model is quite straightforward, and its effectiveness as a useful predictor has been demonstrated recently in the design of an application-specific vocabulary for speech-based data entry into a PDA [3].

Lewis’s model defines ‘true throughput’ (measured in ‘correct words per minute’ - cwpm) as the number of words entered, divided by the time taken to enter them added to the time taken to correct any errors. The basic formula is as follows:-

\[ T = \frac{60 \times R}{60 + (R \times E \times C)} \]

... where \( T \) is the true throughput (in ‘correct words per minute’), \( R \) is the word input entry rate (in ‘words per minute’ - wpm), \( E \) is the word error rate (%) and \( C \) is the time to correct each error (in seconds).

As a simple illustration, the model predicts that an input entry rate of 100 wpm will be reduced to a throughput of under 40 cwpm if the word error rate is 10% and it takes 10 seconds to correct each error.

Since the study outlined in this paper addressed both word-based and character-based methods of data entry, the basic formula was extended to be able to handle character entry and character error rates, and to relate them to throughput in correct words per minute. This required the use of estimates of the average number of characters per word as well as an accommodation of the need to enter the spaces between words explicitly.

4. Calibrating the model

In order to address the alternative input methods discussed in Section 2, the model was set up to reflect typical input entry rates, error rates and correction times based on data derived from the literature and measured from human subjects [2]-[12] (see Table 1). For example, Lewis [2] reports average times to correct an error of 29.1 seconds for speech-based correction and 13.2 seconds for multi-modal correction.

<table>
<thead>
<tr>
<th>INPUT MODE</th>
<th>ENTRY RATE (user type)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dictating to ASR</td>
<td>107 wpm (average)</td>
</tr>
<tr>
<td>Spelling</td>
<td>30 wpm (expert)</td>
</tr>
<tr>
<td>NATO alphabet</td>
<td>17 wpm (expert)</td>
</tr>
<tr>
<td>QWERTY typing</td>
<td>33 wpm (novice) 46 wpm (average) 150 wpm (expert)</td>
</tr>
<tr>
<td>Soft typing [4]</td>
<td>9 wpm (novice) 20 wpm (average) 43 wpm (expert)</td>
</tr>
<tr>
<td>T9® [5]</td>
<td>46 wpm (expert)</td>
</tr>
<tr>
<td>Handwriting [6]</td>
<td>16 wpm (expert)</td>
</tr>
</tbody>
</table>

Table 1: Typical data entry rates for a range of input methods.
The model was also calibrated to accommodate word-based and character-based data entry for both conventional and SMS-style text. A corpus of SMS messages was used in conjunction with an SMS dictionary in order to calculate the average compression ratio for character sequences that could be achieved using SMS-style data entry. Estimates were also obtained for the ratio of ‘normal’ to SMS-style words in typical messages (see Table 2).

<table>
<thead>
<tr>
<th>Average SMS message size</th>
<th>19.58 words (72.11 characters)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average size of a conventional word</td>
<td>5.39 characters</td>
</tr>
<tr>
<td>Average size of an SMS word</td>
<td>3.86 characters</td>
</tr>
<tr>
<td>SMS compression ratio</td>
<td>2.91</td>
</tr>
<tr>
<td>Proportion of normal to SMS words in an SMS message</td>
<td>2.81</td>
</tr>
</tbody>
</table>

*Table 2: Statistics for typical SMS messages.*

5. Results

5.1. Conventional text data entry

Figure 2 illustrates predictions from the model ranked in order of decreasing throughput. For ease of interpretation, only those results relating to ‘expert’ users are presented.

Overall it can be seen that, due to the influence of errors, most input methods achieve a throughput of less than half of their corresponding input data entry rates. This is not the case for the two speech-based spelling modalities, but these have rather low input rates in any case. Also, as suspected, the error rates inherent in contemporary LVCSR are sufficiently high that they lead to dramatic reductions in throughput.

From the figure it can be seen that the best (i.e. fastest) method of data entry is typing on a conventional QWERTY keyboard, with an expert user predicted to achieve a throughput of over 60 cwpm. The worst (i.e. slowest) method is handwriting recognition, with a predicted throughput an order-of-magnitude lower at only 6 cwpm. Other observations that can be made are as follows:

- Standard QWERTY typing is twice as good as its nearest competitors.
- The best speech system - speaker-dependent large vocabulary continuous speech recognition with multi-modal correction (SD LVCSR MMC) - is capable of achieving a throughput that is comparable with soft typing, but the speaker-dependency means that user enrollment is required for the speech-based system.
- The result for speaker-dependent large vocabulary continuous speech recognition with voice-based correction (SD LVCSR VC) successfully predicts the sub-20 cwpm throughput measured in the experiments discussed in Section 1 [1].
- The best hands-free configuration is predicted to be speaker-independent recognition of the spoken conventional orthographic alphabet with voice-based correction (SI Ortho VC) with a throughput of 20 cwpm (slightly slower than soft typing and slightly faster than T9®).
- Despite the low word error rates, speaker-independent recognition of the ICAO phonetic alphabet with voice-based correction (SI ICAO VC) is predicted to be quite slow - 13 cwpm - due to the simple fact that it takes much longer to spell everything out using these special words.
- As expected, T9® is predicted to be faster than multi-press, but both come low in the ranking due to the inherently slower methods of data entry combined with a significant level of input errors.
- Unfortunately, the ideal hands-free system from a user’s perspective - speaker-independent large vocabulary continuous speech recognition with voice-based correction (SI LVCSR VC) - is predicted to give a quite poor throughput of only 10 cwpm.

One possibility that immediately comes to mind is that the predictions for the speech-based methods are based on a speaking rate that, although based on empirical measurement (from dictation to an LVCSR system [4]), is nevertheless somewhat lower than typical conversational rates. In other words, would their ranking be somewhat higher if the users could be encouraged to speak faster? This would, of course, increase the input data entry rate, but the consequence would be that the error rate would be increased and hence the throughput may actually fall.

Somewhat counter-intuitively, the model suggests that it might be better to encourage users to speak more slowly (although contemporary ASR is not necessarily guaranteed to have a better error rate for slow speaking either!).

![Figure 2: Input entry rates and predicted throughput for a range of data entry methods (SD = speaker-dependent, SI = speaker-independent, VC = voice-based correction, MMC = multi-modal correction)](image-url)
5.2. SMS text data entry
The model was also set up to predict the time it would take to enter an average SMS message, either as normal text or using SMS-style contractions. The results are illustrated in Figure 3 ranked in order of increasing data entry time (for normal text).

From Figure 3 it can be seen that, because it is based on the timings for normal text entry, the rankings are the same as in Figure 2. However, clearly a different ranking would be obtained if it were to be based on the data entry times for SMS text. As one would expect, there is no difference between the timings for normal and SMS text for the four LVCSR configurations. This is because whether the text is “later” or “L8R”, the user actually says the same thing.

What can be seen from Figure 3 is that using SMS-style messages gives an improved advantage to all of the character-based data entry methods (such as soft typing and speech-based spelling, for example). Nevertheless, speaker-independent recognition of the spoken conventional orthographic alphabet with voice-based correction (SI Ortho VC) is still not predicted to be faster than speaker-dependent large vocabulary continuous speech recognition with multi-modal correction (SD LVCSR MMC).

Handwriting recognition is predicted to achieve a significant gain from using SMS-style text, but it is still the slowest method of data entry.

6. Conclusions
This paper has described a model for predicting the relative success of alternative method of data entry in terms of the effective ‘throughput’ that is achievable taking into account typical input data entry rates, error rates and error correction times. Results have been presented for the entry of both conventional and SMS-style text.

With regard to the question raised in Section 2 as to whether the higher accuracy obtainable using spelling could mitigate the fact that spelling is inherently slower than speaking normally to an LVCSR system, the results in Figure 2 clearly show that speaker-independent recognition of the spoken conventional orthographic alphabet is predicted to offer quite a significant advantage over speaker-independent large vocabulary continuous speech recognition (both with speech-based correction) - with throughput rates of 21 and 10 cwpm respectively.

Also as to whether speech-input offers an advantage over the small keyboards offered by PDAs and mobile phone, the model predicts that several speech-based configurations with both speech and multi-modal correction can indeed be faster than both T9® and multi-press.

Finally, this paper has shown how a quantitative analytic model can be used to replace time-consuming (and hence expensive) user-based trials in order to provide a first-cut estimate of the relative merits of different data entry technologies. It is recommended that further models of this type should be developed to serve the interests of the speech technology R&D community.

7. References