Actual versus Estimated Dictation Accuracy

TR 29.3790
May 10, 2004

James R. Lewis

IBM Pervasive Computing
Boca Raton, Florida

Abstract

Data from previous studies of dictation accuracy were analyzed to clarify the relationship between actual dictation accuracy and user perception of dictation accuracy. The analysis included data from studies of both discrete and continuous dictation with both male and female speakers. The main effect of type of dictation (discrete vs. continuous) was statistically significant, but the main effect of speaker gender and the interaction between type of dictation and speaker gender were not significant. Speakers' estimates of the accuracy they experienced were consistently lower than the actual accuracy, with the magnitude of the difference greater for discrete than continuous dictation. Regression equations (one for each type of dictation) for predicting actual accuracy from estimated accuracy were both statistically significant. These findings have potential value for future evaluations of dictation accuracy.

ITIRC Keywords

Speech dictation accuracy
Estimated dictation accuracy
Accuracy prediction

Contents

Introduction

The heyday of speech dictation systems seems to have passed, but given the difficulty of data input into handheld devices (Lewis, Potosnak, & Magyar, 1997) and the data input potential of dictation input into handheld devices (Commarford & Lewis, 2004), it is likely that speech dictation applications will re-emerge at some time in the future. An important aspect of speech dictation applications is the dictation accuracy.

In the lab, it is possible to determine the exact accuracy that participants in experiments of dictation applications experience. Outside of the lab, it is rare to have this level of precision, but it is common to have reports of estimated accuracy from various sources such as user reports, product marketing, and journalists (Lewis, 2001). There is currently no published data on the relationship between actual and estimated dictation accuracy. Having such data would be of value in interpreting reported estimates of dictation accuracy when the actual accuracy experienced is unknown.

For several years, the Human Factors team in the Speech Business Unit of IBM[1] conducted experiments of speech dictation accuracy and throughput as part of the development of various dictation products, from discrete dictation products such as VoiceType Dictation[2] and Simply Speaking[3] to the continuous dictation ViaVoice[4] products[5]. Participants in those studies estimated the accuracy they experienced during the dictation sessions they completed in the more formal recognition accuracy studies. The purpose of the analyses presented in this report is to investigate the relationship between actual and estimated dictation accuracy using data gathered in previous experiments.

.

[1] IBM is a registered trademark of International Business Machines Corp.

[2] VoiceType and VoiceType Dictation are trademarks or registered trademarks of International Business Machines Corp.

[3] Simply Speaking is a trademark or registered trademark of International Business Machines Corp.

[4] ViaVoice is a trademark or registered trademark of International Business Machines Corp.

[5] Discrete dictation refers to dictation applications in which participants must pause between each dictated word. In continuous dictation application, no such pause is necessary.

Method

Across the data from the available experiments were 63 independent cases in which each participant provided an estimated accuracy following a dictation test session, with 35 discrete and 28 continuous cases. The data allowed the evaluation of two variables -- whether the recognition system was discrete or continuous, and whether the speaker was male or female.

Results

Analysis of Variance

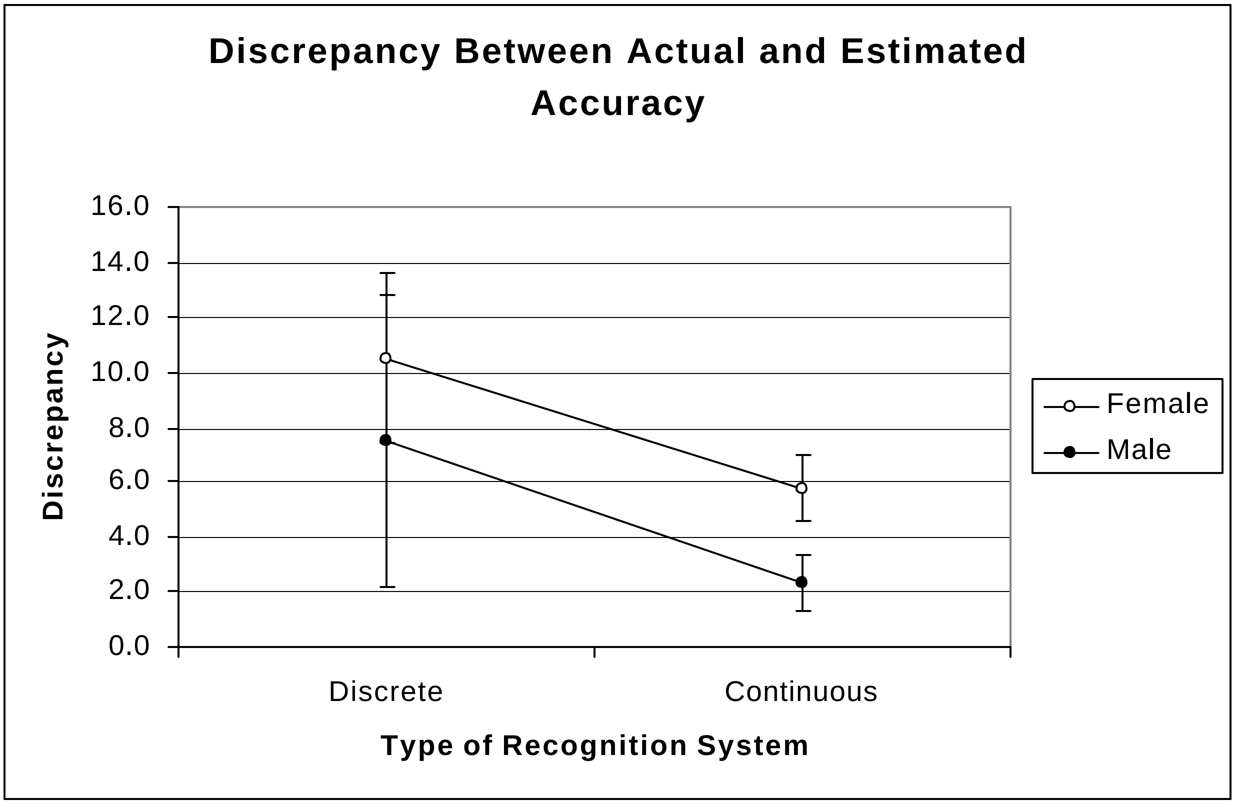
An analysis of variance on the data showed that the type of recognition system had a significant effect on the discrepancy between actual and estimated accuracy (*F*(1,59)=4.9, *p*=.03), but the effect of gender (*F*(1,59)=2.1, *p*=.15) and the interaction between type of system and gender (*F*(1,59)=.01, *p*=.93) were not significant.

The following table and figure show the means and 90% confidence interval deltas for the data (difference scores calculated by subtracting the estimated accuracy from the actual accuracy experienced by the user) broken down by type of system and gender of speaker.

Table 1. Means and 90% Confidence Interval Deltas

	Means		90% CI Deltas	
	Discrete	Continuous	Discrete	Continuous
Female	10.5	5.7	3.1	1.2
Male	7.5	2.3	5.3	1.0

Figure 1. Discrepancy Between Actual and Estimated Accuracy



Because the difference scores were calculated by subtracting the estimated accuracy from the actual accuracy, the positive results show that users strongly tended to provide an estimate of accuracy that was lower than the accuracy they actually experienced. Each of the means in the table is significantly different from 0 (see Table 2 for the applicable *t*-tests).

Table 2. t-tests on Tabled Values (testing hypothesis that tabled value is really 0)

Discrete Female:	t(16)=5.6, p<.00001
Continuous Female:	t(13)=7.8, p<.00001
Discrete Male:	t(17)=2.3, p=.03
Continuous Male:	t(13)=3.7, p=.002

Regression Analyses

Table 3 shows the results of regressio n analyses. Because the main effect of dictation type (discrete vs. continuous) was significant, I developed a model for each type.

Table 3. Regression Analyses

```
THE FOLLOWING RESULTS ARE FOR:
             TYPE$    = Continuous

DEP VAR:ACCURACY      N:      28  MULTIPLE R: 0.758  SQUARED MULTIPLE R: 0.575
ADJUSTED SQUARED MULTIPLE R:  .559    STANDARD ERROR OF ESTIMATE:        2.797

VARIABLE      COEFFICIENT    STD ERROR     STD COEF TOLERANCE    T   P(2 TAIL)

CONSTANT           48.112        6.821        0.000      .       7.053    0.000
ESTACC              0.478        0.081        0.758     1.000    5.934    0.000

                       ANALYSIS OF VARIANCE

SOURCE       SUM-OF-SQUARES   DF  MEAN-SQUARE     F-RATIO       P

REGRESSION         275.368     1      275.368      35.207       0.000
RESIDUAL           203.356    26        7.821


THE FOLLOWING RESULTS ARE FOR:
             TYPE$    = Discrete

DEP VAR:ACCURACY      N:      35  MULTIPLE R: 0.635  SQUARED MULTIPLE R: 0.404
ADJUSTED SQUARED MULTIPLE R:  .386    STANDARD ERROR OF ESTIMATE:        4.535

VARIABLE      COEFFICIENT    STD ERROR     STD COEF TOLERANCE    T   P(2 TAIL)

CONSTANT           67.506        4.497        0.000      .      15.012    0.000
ESTACC              0.264        0.056        0.635     1.000    4.727    0.000

                       ANALYSIS OF VARIANCE

SOURCE       SUM-OF-SQUARES   DF  MEAN-SQUARE     F-RATIO       P

REGRESSION         459.544     1      459.544      22.347       0.000
RESIDUAL           678.623    33       20.564
```

From the results shown in Table 3, the formula for predicting actual discrete dictation accuracy from estimated discrete accuracy is ACCURACY = 67.506 + .264*ESTIMATED. The variability in ESTIMATED accounted for about 40% of the variability in ACCURACY. Both the constant and slope for the regression equation were highly significant ($p < .0001$). The 95% confidence delta for predictions with the equation was +/- 9%.

The formula for predicting actual continuous dictation accuracy from estimated continuous accuracy is ACCURACY = 48.112 + .478*ESTIMATED. The variability in ESTIMATED accounted for about 58% of the variability in ACCURACY. Both the constant and slope for the regression equation were highly significant ($p < .0001$). The 95% confidence delta for predictions with the equation was +/- 5.6%.

Discussion

Even though dictation accuracy data for handheld devices is not yet available, it is likely that at some time in the future there will be published data regarding both actual and estimated dictation accuracy with these types of devices. Over time it will be possible to collect handheld data similar to the data presented in this report, but it will take time. The data presented in this report will provide a reasonable approximation to use until handheld data becomes available.

It is important to keep in mind that this data was collected during the course of testing software, and there might be additional effects that would make this data less reliable in the marketplace. For example, it is possible that a user who purchases the software and experiences accuracy lower than he or she expected might experience anger that would lead them to estimate their accuracy as poorer than the people who participated in these experiments. Despite this, the evidence that the participants provided lower estimates of accuracy than they actually experienced is strong, and this should be taken into account when interpreting estimates of dictation accuracy for which there is no corresponding measurement of the actual dictation accuracy.

References

Commarford, P. M., & Lewis, J. R. (2004). Models of throughput rates for dictation and voice spelling for handheld devices. *International Journal of Speech Technology*, *7*, 69-79.

Lewis, J. R. (2001). *The accuracy wars: Journalists' estimates of continuous speech product dictation accuracy from 1997-1999* (Tech. Report 29.3465). Raleigh, NC: IBM Corp.

Lewis, J. R., Potosnak, K. M., & Magyar, R. (1997). Keys and keyboards. In M. Helander, T. K. Landauer, and P. V. Prabhu (Eds.), Handbook *of Human-Computer Interaction* (pp. 1285-1315). Amsterdam: North-Holland.