Technology Evaluation ■

Comparative Evaluation of Three Continuous Speech Recognition Software Packages in the Generation of Medical Reports

ERIC G. DEVINE, PHD, STEPHAN A. GAEHDE, MD, MPH, ARTHUR C. CURTIS, MD, PHD

Abstract **Objective:** To compare out-of-box performance of three commercially available continuous speech recognition software packages: IBM ViaVoice 98 with General Medicine Vocabulary; Dragon Systems NaturallySpeaking Medical Suite, version 3.0; and L&H Voice Xpress for Medicine, General Medicine Edition, version 1.2.

Design: Twelve physicians completed minimal training with each software package and then dictated a medical progress note and discharge summary drawn from actual records.

Measurements: Errors in recognition of medical vocabulary, medical abbreviations, and general English vocabulary were compared across packages using a rigorous, standardized approach to scoring.

Results: The IBM software was found to have the lowest mean error rate for vocabulary recognition (7.0 to 9.1 percent) followed by the L&H software (13.4 to 15.1 percent) and then Dragon software (14.1 to 15.2 percent). The IBM software was found to perform better than both the Dragon and the L&H software in the recognition of general English vocabulary and medical abbreviations.

Conclusion: This study is one of a few attempts at a robust evaluation of the performance of continuous speech recognition software. Results of this study suggest that with minimal training, the IBM software outperforms the other products in the domain of general medicine; however, results may vary with domain. Additional training is likely to improve the out-of-box performance of all three products. Although the IBM software was found to have the lowest overall error rate, successive generations of speech recognition software are likely to surpass the accuracy rates found in this investigation.

■ **J Am Med Inform Assoc.** 2000;7:462–468.

Changes in health care are increasing the demand for electronic records in large organizations. Medical professionals who do not have access to transcription services must type their own chart entries, which requires typing skill and significant amounts of time. Because of increased sharing of patient care across multiple facilities, VA New England is interested in evolving technology-based approaches to enhancing documentation of patient care in an electronic form. This study was undertaken in part to assess the potential use of speech recognition software in busy clinical settings without transcription support, prior to a decision on significant capital investment.

By the close of the 1990s, speech recognition software had become a potentially viable and affordable substitute for transcription, costing approximately $2,000 per workstation with software. Software that converts

Affiliation of the authors: Boston Veterans Administration Medical Center, Boston, Massachusetts.

This work was supported in part by the Veterans Affairs Medical Informatics Fellowship Training Program.

Correspondence and reprints: Eric Devine, PhD, Department of Psychology (116b), Boston Department of Veterans Affairs, 150 S. Huntington Avenue, Boston, MA 02130;
e-mail: ⟨devine.eric@boston.va.gov⟩.

Received for publication: 1/18/00; accepted for publication: 5/18/00.

the spoken word to text has been used in many specialized health care settings (e.g., radiology and cardiology). A search of medical and psychological journal listings (using MEDLINE and PsychLit) revealed few published articles evaluating speech recognition software in health care settings. It is noteworthy that the majority of these studies evaluated discrete speech recognition software. A number of software reviews have been published in the popular press and computer trade magazines (e.g., *PC Magazine*, Nov 1999), but none of these publications has provided a systematic comparison of continuous speech recognition software package performance.

Zafar et al.[1] conducted the only published study evaluating continuous speech recognition software. Their article includes a comprehensive overview of the process of continuous speech recognition, a comparative evaluation of the specifications of three different products (IBM ViaVoice Gold, Dragon Systems NaturallySpeaking, and Philips SpeechMagic), and findings from tests conducted with these three products. The authors reported accuracy rates as high as 98 percent with significant training of the software, but their methods for categorizing and counting recognition errors were not described. Although the authors tested three different software packages, they did not present comparative data that would allow evaluation of the relative accuracy of each software package. In addition, it was unclear from the results whether the recognition rate reported was for a single dictator or was the average of the three authors' recognition rates. Furthermore, the high rate of recognition may be related to the skill of the small number of dictators who evaluated the software.

The primary objective of the present study was to compare the relative accuracy of three continuous speech recognition software packages. The study was designed to evaluate out-of-box performance that may be expected, with minimal training of the software, by both experienced and inexperienced dictators. In addition, the study was designed to allow for comparative evaluation of error rates within specific categories of errors (e.g., medical abbreviations and punctuation).

Materials and Methods

Materials

Software

The following continuous speech recognition packages were evaluated in this study: IBM ViaVoice 98 with IBM General Medicine vocabulary (IBM, Armonk, New York); Dragon NaturallySpeaking Medical Suite, version 3.0 (Dragon Systems, Inc., Newton, Massachusetts); and L&H Voice Xpress for Medicine, General Medicine Edition, version 1.2 (Lernout & Hauspie, Burlington, Massachusetts).

Hardware

Pentium II computers of identical specifications were used for comparison of the three software packages. Each computer was a 333-MHz system equipped with 128 MB of RAM and a Creative SoundBlaster–compatible sound card.

Subjects

Twelve physicians working with the VA New England Healthcare System participated in the study. Participants were all male (an accident of availability) and ranged in age from 29 to 59 years, with a mean age of 46 years. The majority (11 of 12) spoke English as a native language. All participants reported using a computer at least two hours a day, and almost half (5 of 12) reported previous experience with voice dictation software. Only two participants reported ongoing use of speech recognition software at the time of the study.

Medical Record Samples

Four sample medical record entries (two progress notes, one assessment summary, and one discharge summary) were chosen for dictation based on vocabulary, formatting complexity, and length. All material was drawn from actual records. Identifying information contained in the reports was altered to preserve patient confidentiality, and the reports were edited for typographic errors.

Procedure

All testing of speech recognition software was conducted in a single day for each clinician. Each clinician completed enrollment and dictation trials for the three software packages. Because of the potential threats to validity posed by a within-subjects design (e.g., fatigue, learning effects), the order of use for each software package was counterbalanced using a Latin square design.

Enrollment

Each clinician completed the standard voice enrollment for each software package. Although additional training of the speech model is available in each package, and extended training of voice recognition software will improve performance and ease of use, an abbreviated training period (less than 60 min) was chosen in this study to evaluate this software for use in a setting in which extensive training may not be

Table 1 ■

Voice Scoring Categories

General English vocabulary:
General vocabulary word omitted
Word misrecognized as different word
Word misrecognized as two or more different words
Homophone substituted
General vocabulary recognized as medical vocabulary
Proper name misrecognized
General vocabulary abbreviation misrecognized
Medical vocabulary:
Medical vocabulary recognized as general vocabulary
Medical vocabulary recognized as wrong medical vocabulary
Medical vocabulary recognized as two or more words
Medical abbreviation recognized as general vocabulary
Medical abbreviation recognized as wrong medical abbreviation
Medical abbreviation recognized as two or more words
Medical abbreviation omitted
Extra word omitted
Numbers:
Number misrecognized as general vocabulary
Numbers omitted
Wrong number recognized
Punctuation:
Punctuation mark omitted/spelled out/recognized as general text:
Period
Comma
Colon
Semicolon
Quotes
Dash
Parentheses
Slash
Wrong punctuation recognized

practically feasible. Following voice enrollment, participants completed two medical chart dictations (a medical progress note, and one psychological report) for practice and then began the dictation trials.

Dictation Trials

For the scored trials, each clinician completed a 707-word medical discharge summary and a 257-word medical progress note. The progress note contained 227 words of general English vocabulary, 17 words of medical vocabulary, 13 medical abbreviations, 10 numbers, and 48 punctuation marks. The discharge summary contained 568 words of general English vocabulary, 98 words of medical vocabulary, 41 medical abbreviations, 90 numbers, and 216 punctuation marks. Medical vocabulary was defined as words that were unlikely to appear outside a medical context (e.g., erythematous); English vocabulary was defined as words that can be found in nonmedical prose (e.g., trauma). Dictators were instructed not to correct errors in dictation either by voice or by typing. Time to complete each dictation was recorded, and a copy of the generated file was saved for evaluation of speech recognition errors. Participants provided a subjective rating of each software package following completion of the dictation trial with each package and then again following completion of trials with all three packages.

Dictation Sample Scoring

Errors in dictation were assessed by word-for-word comparison of a printed copy of the dictation samples and the captured dictations. Errors in recognition were categorized and recorded in 43 distinct groups, which cluster into five broad areas: general English vocabulary misrecognition, medical vocabulary misrecognition, extra word insertion, number misrecognition, and punctuation misrecognition. (Table 1 shows a breakdown of each error type within the broad categories.) Several scoring procedures were implemented to improve the consistency of scoring: 1) all dictations were scored in a group (of three investigators and two research assistants) so that all scoring issues could be decided by group consensus; 2) each dictation (across packages) for a single participant was scored by the same investigator; and 3) after all dictation samples for the medical progress note had been scored, investigators performing all subsequent scoring were blinded to the software used. Rules were also implemented to ensure consistency of scoring. For example, the maximum number of possible errors in a dictation was based on the number of items in the dictation sample, and parts of speech that improved grammar but were not present in the dictation sample were not counted as errors. In addition, many minor rules (not described in the present paper) specific to each error category were used to guide scoring (e.g., ''low back pain'' is an acceptable substitute for ''LBP'').

Results

The study was designed to evaluate the accuracy of three continuous speech recognition products in the generation of medical chart entries. Examination of combined errors across categories (general English vocabulary, medical vocabulary, medical abbreviations, numbers, and punctuation) revealed that the IBM software had the lowest mean error rate (6.6 to 8.4 percent) followed by the Dragon software (12.0 to 13.9 percent) and then the L&H software (13.8 to 14.6 percent). Examination of the overall error rate for recognition of vocabulary alone (General English vocabulary, medical vocabulary, medical abbreviations) yielded similar results. IBM had the lowest mean error rate (7.0 to 9.1 percent) followed by L&H (13.4 to 15.1 percent) and then Dragon (14.1 to 15.2 percent).

Order Effects

Repeated measures analyses of variance were conducted to test for any order effects that may have re-

sulted from the within-subjects design. Order of use for the three software packages was entered as a between-subjects variable, and total error scores for each package were entered as the within-subjects variables. Analysis of the progress note data revealed no significant effect for order of use ($F[2,8] = 0.013$, $P < 0.987$). Similar results were found with the discharge summary data ($F[2,9] = 0.248$, $P < 0.785$).

Time to Complete Dictation Trials

The length of time needed to dictate the 257-word progress note was consistent across products, with no significant differences ($F[2,22] = 1.05$, $P < 0.336$) in the mean dictation time for IBM ($M = 6.0$ min), Dragon ($M = 5.2$ min), and L&H ($M = 6.4$ min) software. The length of time needed to complete a 938-word discharge report, however, was significantly different between packages ($F[1,10] = 10.9$, $P < 0.01$], with Dragon taking the shortest amount of time ($M = 12.2$ min), followed by IBM ($M = 14.7$ min), and L&H ($M = 16.1$ min).

Comparative Error Rates Across Packages

Analyses of variance of the discharge summary data revealed significant differences in error rates among software packages in the recognition of general English vocabulary, medical vocabulary, medical abbreviations, numbers, and punctuation. (Table 2 shows ANOVA results.) Slightly different findings emerged in the analyses of the progress note data. Significant differences in error rates were found in the recognition of general English vocabulary, medical abbreviations, and numbers but not in recognition of medical vocabulary and punctuation.

Overall Speech Recognition Error Rates

To examine the overall error rates for the three packages, the ratio of errors (observed errors/possible errors) was examined for each category in which a significant difference was found among packages. Table 3 shows that for the progress note, IBM had the lowest rate of errors for general English vocabulary ($M = 7.65$ percent), medical abbreviations ($M = 23.78$ percent), and numbers ($M = 13.64$ percent). The overall error rate for all items combined (words plus numbers and punctuation) was lowest for IBM (8.40 percent) followed by L&H (13.85 percent) and Dragon (13.88 percent). A series of pairwise comparisons showed that the IBM error rate was significantly different from both the Dragon error rate and the L&H error rate for both English vocabulary and medical abbreviations. The error rate for recognition of numbers was significantly different between IBM and L&H and also between Dragon and L&H, but not between IBM and Dragon.

Table 2 ■

ANOVA Results for the Progress Note and Discharge Summary

	DF	*F*	*P<*
Discharge summary:			
General English vocabulary	2,22	11.54	0.000
Medical vocabulary	2,22	11.77	0.000
Medical abbreviations	2,22	27.26	0.000
Numbers	2,22	17.06	0.000
Punctuation	2,22	12.43	0.000
Progress note:			
General English vocabulary	2,20	4.70	0.020
Medical vocabulary	2,20	0.52	0.594
Medical abbreviations	2,20	5.15	0.016
Numbers	2,20	6.97	0.005
Punctuation	2,20	2.71	0.091

Similar results were found for the discharge summary. IBM had the lowest rate of errors for general English vocabulary ($M = 6.22$ percent), medical vocabulary ($M = 9.10$ percent), medical abbreviations ($M = 13.01$ percent), numbers ($M = 10.56$ percent), and punctuation ($M = 3.70$ percent). Table 4 shows that the overall error rate (all items combined) was lowest for IBM (6.62 percent) followed by Dragon (12.03 percent) and L&H (14.62 percent). A series of pairwise comparisons showed that the IBM error rate was significantly different from the L&H error rate for all error categories. Significant differences between IBM and Dragon emerged for English vocabulary, medical vocabulary, and medical abbreviations, but not for punctuation and numbers. Significant differences between L&H and Dragon were found for both medical abbreviations and numbers.

Overall Recognition Rates by Previous Dictation Experience

A series of independent sample t-tests were conducted to determine whether participants who had significant experience with dictation (at least two years of experience with either a transcription service or voice recognition software) achieved better rates of correct recognition than participants who had no experience with dictation. Analysis of the progress note data showed that past dictation experience was not related to the performance of the IBM package ($t = 0.952$, $P < 0.366$), the Dragon package ($t = 1.80$, $P < 0.105$), or the L&H package ($t = 1.53$, $P < 0.366$). Analysis of the discharge data revealed similar results.

Comparative Evaluation of Findings

A series of paired sample t-tests were conducted to determine whether there was a significant difference

Table 3 ■

Mean Error Rates and Percentage of Misrecognized Words in Each Category and Across Products for the Progress Note Dictation

		Mean (%) Errors for Each Product		
	No. Items	IBM	Dragon	L&H
General vocabulary	227	17.36 (7.65)	29.09 (12.87)	26.64 (11.79)
Medical vocabulary	17	3.00 (17.65)	3.91 (22.39)	3.09 (18.18)
Medical abbreviations	13	3.09 (23.78)	6.18 (48.00)	4.82 (37.06)
Numbers	10	1.36 (13.64)	2.00 (20.00)	4.00 (40.00)
Punctuation marks	48	1.64 (3.41)	2.55 (5.30)	3.55 (7.39)
TOTAL	315*	26.45 (8.4)	43.73 (13.88)	43.64 (13.85)

*The total number of items reflects the 257 words plus 10 numbers and 48 punctuation marks.

Table 4 ■

Mean Error Rates and Percentage of Misrecognized Words in Each Category and Across Products for the Discharge Summary Dictation

		Mean (%) Errors for Each Product		
	No. Items	IBM	Dragon	L&H
General vocabulary	568	35.33 (6.22)	70.00 (12.70)	65.58 (11.90)
Medical vocabulary	98	8.92 (9.10)	15.42 (15.73)	19.00 (19.39)
Medical abbreviations	41	5.33 (13.01)	14.08 (34.35)	22.17 (54.07)
Numbers	90	9.50 (10.56)	12.00 (13.33)	19.42 (21.57)
Punctuation marks	216	8.00 (3.70)	10.33 (4.78)	21.92 (10.15)
TOTAL	1,013*	67.08 (6.62)	121.83 (12.03)	148.08 (14.62)

*The total number of items reflects the 707 words plus 90 numbers and 216 punctuation marks.

between mean error rates for each package between dictation trials. Mean percent error rates for the discharge summary and progress note data were not found to be statistically different for IBM ($t = -0.677$, $P < 0.512$), Dragon ($t = -0.354$, $P < 0.730$), and L&H ($t = 0.964$, $P < 0.356$) software packages.

Subjective Ratings

In response to the question ''Would you use this software again?'' 100 percent of participants replied ''Yes'' for both the IBM and Dragon products, whereas only 66 percent of participants replied ''Yes'' for the L&H product. Following use of all three products, participants were asked to rank order each system on the basis of their perception of the product 's performance. The IBM product received the most favorable responses, with 92 percent (11 of 12 participants) ranking it number one, and 8 percent (1 of 12 participants) ranking it number two. The Dragon product was rated number one by 17 percent (2 of 12 participants) and number two by 83 percent (10 of 12 participants),* whereas the L&H product was ranked number three by 100 percent (12 of 12 participants).

Discussion

Results of this study suggest that, in generating medical record entries, the out-of-box performance of IBM ViaVoice 98 is better than that of software developed by Dragon and L&H; however, recognition rates may vary, depending on speech domain. Specifically, across both dictation samples, the IBM product was superior to the other two products in the recognition of general English vocabulary, medical abbreviations, and numbers. Findings for medical vocabulary and punctuation were not consistent across test dictation trials, so conclusions about the relative performance

*One participant ranked the IBM and Dragon products as ''tied'' for number one and rated the L&H product number three.

of the three packages for these two error categories are limited.

Surprisingly, previous experience with dictation services did not have an effect on the overall error rates. This result may be due partly to two factors: 1) participants had some time to practice dictation of the two practice notes prior to the scored dictation trials, and 2) participants were reading from a script and thus did not experience some of the slowing and stopping that is typical of inexperienced dictators as they learn to compose notes as they dictate. This latter point may have also had some effect on recognition rates across the three packages. Specifically, recognition rates may have been enhanced by the absence of pauses and hesitations that would have been present if the dictators were required to compose medical entries instead of reading from a script.

Also notable was the high rate of recognition errors found in the use of technology that is more advanced than discrete speech recognition. Error rates reported in studies examining discrete speech recognition software have ranged from 1 to 3 percent, which is significantly lower than the 6 to 8 percent reported in the present study for the most accurate of the three packages evaluated. Based on the results of Zafar et al.,[1] however, it is reasonable to conclude that extended training of the speech model, extended training of the dictator in package-specific dictation conventions, and the addition of vocabulary not contained in products as shipped would improve the recognition rates substantially.

The first methodologic challenge of this study was to develop a scoring protocol that would ensure the consistency of scoring across packages within a subject and across subjects within the sample. The primary strategy used in developing this protocol was to limit the degree of interpretation that was needed in scoring an error. This strategy was adopted because interpretive scoring protocols are difficult to standardize and are threatened by any potential interpretive biases. In pursuit of this strategy, we developed some simple rules for scoring, which did not allow the evaluator to score dictations on the basis of an interpretation of what the evaluator thought went wrong in the dictation. Although a rigid application of these rules removed the guesswork and potential errors in scoring, these rules did not allow us to develop a better understanding of the types of errors that may be inherent in the speech models of the software evaluated, and they may have contributed to higher absolute error rates.

Some methodologic challenges were difficult to address because of time and cost constraints. In this study design, for example, we were unable to determine whether words missing from the dictation sample were missing because they were misrecognized by the software or because the dictator omitted them. In fact, in several dictation samples, it was clear that the dictator had omitted a word or phase in preference for some other style of presenting the information. Scoring of these instances, however, was not altered to take account of the dictator ‘s missed words, as our scoring methodology was designed to limit the degree of interpretive bias that might be present. With greater resources, it might have been possible to audiotape the dictators ‘ speech, have that audiotape transcribed to a typed report, and evaluate the recognition errors for the dictation sample on the basis of the transcribed report. Having 64 dictation samples transcribed, however, was beyond the scope of our budget. Although this may seem to be a serious source of error in our findings, the study was designed so that the occurrence of this type of error should be distributed evenly across software packages.

With advances in technology since the time of this study, several improvements in the present methodology have become possible. Future studies should make use of digital voice recorders that are designed to interface with speech recognition software. Use of this technology would allow each participant in the study to dictate each report only once and then use the same speech file with each software package. This would eliminate many potential confounders of a within-subjects design (e.g., practice effects, fatigue) and it would also be more economically feasible to have the digital speech file transcribed for direct comparison with the dictation sample.

Conclusions

With increasing power, decreasing cost of computing hardware, and recent increases in the sophistication of speech recognition software, the use of speech recognition to replace transcription in real-world settings is finally becoming feasible. Accuracy of recognition, which has traditionally been a major problem, has increased dramatically because of improvements in speech recognition technology and availability of medical vocabularies. As a result, it is becoming possible to think of productively employing off-the-shelf speech recognition products in clinical settings.

However, a thorough understanding of the business problems that need to be solved and the technical and functional attributes of available products is essential to any successful undertaking. The study discussed

here is important because it represents one of the veryfew attempts at a robust evaluation of performance attributes of commercial speech recognition products. Although technology will continue to evolve, and products will come and go, knowledge of the approach taken in this study should be of value both to health care organizations considering speech recognition implementations and to researchers contemplating further investigation of speech recognition technology.

Reference ■

1. Zafar A, Overhage M, McDonald CJ. Continuous speech recognition for clinicians. J Am Med Inform Assoc. 1999;6(3): 195–204.