# Initial Evaluation of a Continuous Speech Recognition Program for Radiology

Kalpana M. Kanal, Nicholas J. Hangiandreou, Anne-Marie G. Sykes, Heidi E. Eklund, Philip A. Araoz, Jorge A. Leon, and Bradley J. Erickson

**The aims of this work were to measure the accuracy of one continuous speech recognition product and dependence on the speaker's gender and status as a native or nonnative English speaker, and evaluate the product's potential for routine use in transcribing radiology reports. IBM MedSpeak/Radiology software, version 1.1 was evaluated by 6 speakers. Two were nonnative English speakers, and 3 were men. Each speaker dictated a set of 12 reports. The reports included neurologic and body imaging examinations performed with 6 different modalities. The dictated and original report texts were compared, and error rates for overall, significant, and subtle significant errors were computed. Error rate dependence on modality, native English speaker status, and gender were evaluated by performing *t* tests. The overall error rate was 10.3 ± 3.3%. No difference in accuracy between men and women was found; however, significant differences were seen for overall and significant errors when comparing native and nonnative English speakers ($P = .009$ and $P = .008$, respectively). The speech recognition software is approximately 90% accurate, and while practical implementation issues (rather than accuracy) currently limit routine use of this product throughout a radiology practice, application in niche areas such as the emergency room currently is being pursued. This methodology provides a convenient way to compare the initial accuracy of different speech recognition products, and changes in accuracy over time, in a detailed and sensitive manner.**


**KEY WORDS: continuous speech recognition, radiology transcription, computers.**

RADIOLOGY REPORTS in most medical settings are generated according to the following steps. As the images for each examination are read by the radiologist, a verbal report is dictated and recorded. This audio report is typed by a human transcriptionist, resulting in a text report. The text report is considered preliminary, but in some practices it may be shared with clinicians at this stage. The radiologist then finalizes the transcribed report after reviewing it (usually without reviewing the images) and assuring the accuracy of the text. The final report and images are then made available to clinicians. Time delays between the various stages of this process usually mean that final reports (and even preliminary reports) are available only after several hours or more have passed after examination interpretation.

In some high-volume areas in the Mayo Clinic (Rochester, MN) radiology practice, a transcriptionist is present in the reading room and types the report as it is dictated by the radiologist. The text report is then reviewed quickly for accuracy by the radiologist while the images are still available, before going on to the next case. In this way, the *final* report is produced immediately after examination interpretation, and the turnaround time of this final report to the clinician is improved significantly. However, many areas of this practice do not produce a large enough examination volume to justify a dedicated transcriptionist in the reading room, and these areas suffer from increased report turnaround time compared with areas in which direct dictation is practiced.

The emergence of automatic speech recognition software has suggested that all reading rooms might operate in the direct dictation mode. Especially when used in conjunction with electronic systems for managing the text information (Radiology Information System, or RIS) and image information (Picture Archiving and Communication System, or PACS), speech recognition software may allow all finalized radiology examinations, consisting of the report and the images, to be delivered to clinicians with minutes of interpretation by the radiologist. This immediate delivery of the radiology product to the primary care physicians would be expected to have a significant positive impact on overall quality of patient care in many areas of a medical practice.

Early speech recognition software products required the user to speak in a discontinuous manner, so that each individual word could be identified and transcribed.[1-11] Overall accuracy rates in word

*From the Department of Radiology, University of Washington, Seattle, WA; and the Department of Radiology, Mayo Clinic, Rochester, MN.*

*Address reprint requests to Kalpana M. Kanal, PhD, University of Washington, Department of Radiology, Box 357115, 1959 NE Pacific St, Seattle, WA 98195-7115.*



*0897-1889/01/1401-0006$35.00/0*

*doi:10.1053/jdim.2001.21683*

recognition using these discrete speech systems have been reported to be as high as approximately 98%.[3,9] The requirement for discrete or discontinuous speech made early speech recognition products impractical for routine use in a typical high-volume radiology reading room in spite of their high accuracy rates. Newer products allow the user to speak in a more natural, continuous manner.[12-18] Zafar et al,[17] Schwartz et al,[19] and Bergeron[20] list available speech recognition systems and product characteristics.

Our aims in the current work include measurement of the initial accuracy of 1 continuous speech recognition product, investigation of the impact on accuracy of the sex of the speaker and status of the speaker as a native or nonnative English speaker, and evaluation of the potential for routine clinical use of the system for radiology report transcription. In addition, we wish to establish the validity of our methodology as an efficient tool for comparing different speech recognition systems within an institution based on computing error rates.

## MATERIALS AND METHODS

### *Speech Recognition Software and Computer System*

IBM MedSpeak/Radiology Software, version 1.1 (Health Care Industry Solutions Unit; IBM Corporation, Hawthorne, NY) is the application that was evaluated. This software allows continuous speech to be transcribed to text as it is spoken. The speech recognition software was run on a Pentium Pro 200 MHz/256K cache personal computer (PC) system (Dell Computer Corporation, One Dell Way, Round Rock, TX). The system has 128 megabytes (MB) of RAM, 2 gigabytes (GB) of hard disk space, and was equipped with a SoundBlaster 16 sound board (recommended by IBM for use with the MedSpeak/Radiology software), and Altec speakers. The total computer system cost was under $4,500 (when originally purchased). The software package shipped by IBM included a microphone, which was used in our tests.

### *Speakers, Enrollment, and Dictation*

Six speakers, 3 men and 3 women, participated in the study. All speakers were from the radiology department and were familiar with medical and radiology terminology. Two of the speakers were medical physicists, and the remaining 4 speakers were radiology residents. The group of 6 speakers was divided evenly according to gender, and 2 of the 6 participants were nonnative English speakers (1 man and 1 woman). These speakers were selected to reasonably represent our clinical practice, which consists of a diverse group of radiologists. This also limits any bias from being introduced into our selection of speakers. If the speakers chosen had been limited to a "white only" or "men only" for example, a definite bias would have been introduced into our data. It should be noted that there is a wide spectrum of proficiency among nonnative speakers across institutions. However, this experimental setup is designed for use within 1 institution to compare various speech recognition products or compare 1 product over a period of time, using the same representative set of speakers. Each speaker performed the optional enrollment procedure, which, according to the software vendor, trains the recognition software to respond to an individuals voice and speech patterns with greater accuracy. Each speaker reads 50 sentences displayed by the program to complete the minimum enrollment process, which required approximately 15 to 20 minutes. The system allows a maximum of 200 sentences for enrollment, and 1 speaker (speaker A) enrolled using both the 50 and 200 sentence sets. All software and computer settings were kept constant during the entire enrollment process as well as during the actual dictation of the reports.

A test set of 2 reports from each of 6 imaging modalities (conventional x-ray, ultrasound scan, mammography, nuclear medicine, magnetic resonance imaging [MRI], and computed tomography [CT] were selected randomly for use in the study, resulting in a total of 12 reports in the test set. Reports from both body and neurologic imaging examinations were represented in the test set. After enrollment, each speaker dictated the 12 reports in random order. Only the main body of the report text was dictated (omitting recitation of the patient name, identification number, and indication or diagnosis codes). Dictation was carried out in a quiet radiology reading room. The speakers were advised to speak using their normal rate and volume, and to avoid watching the computer monitor while dictating. The microphone position varied somewhat with each reader according to individual preference. We did not control this because in a practical clinical environment, this is totally dependent on the individual user. Our experimental setup was designed to closely imitate a realistic clinical setting. However, it should be noted that during the enrollment process, it was observed that the speech recognition system acknowledged reasonable recognition accuracy during an individual speaker's enrollment. Dictation of the test set of reports required approximately 12 to 15 minutes. No corrections were made to the dictated reports produced by the software. These dictated reports were saved as text files on the computer hard disk for later comparison with the text of the original report.

### *Report Evaluation and Error Classification*

Each report transcribed by the speech recognition software was compared automatically with the original report text using a software program for text document comparison (DocuComp II version 1.0; Advanced Software, Inc, Sunnyvale, CA). This software program is 100% accurate in identifying discrepancies and can be used to compare any document, not necessarily those of a medical nature. However, 2 of the authors reviewed the compared documents to ensure that no discrepancies were omitted. Once the original and dictated reports were compared, each discrepancy was classified as 1 of 4 different error types. Three authors determined the classification of errors. Class 0 errors involved no change in meaning with respect to the original report text, and the transcribed text was grammatically correct. Class 1 errors also involved no change in meaning, but the transcribed text grammatically was incorrect. Class 2 errors were those in which the meaning of the transcribed report text was different than that of the original report text, but the error

**Table 1. Classification of Errors in Each Transcribed Report for Speaker A**

| Report | Number of Errors Observed | | | | Words in Original Report |
|---|---|---|---|---|---|
| | Class 0 | Class 1 | Class 2 | Class 3 | |
| CT 1 | 0 | 6 | 5 | 2 | 207 |
| CT 2 | 0 | 1 | 8 | 1 | 68 |
| Mammo 1 | 0 | 2 | 2 | 0 | 66 |
| Mammo 2 | 0 | 2 | 1 | 1 | 58 |
| MRI 1 | 2 | 2 | 2 | 0 | 61 |
| MRI 2 | 0 | 3 | 8 | 0 | 126 |
| Nuc Med 1 | 0 | 3 | 4 | 1 | 67 |
| Nuc Med 2 | 0 | 4 | 6 | 0 | 130 |
| US 1 | 0 | 1 | 5 | 2 | 107 |
| US 2 | 0 | 2 | 3 | 0 | 63 |
| X-ray 1 | 0 | 1 | 0 | 0 | 20 |
| X-ray 2 | 0 | 2 | 0 | 0 | 67 |

NOTE. Similar error classification also was performed for the remaining 5 speakers.

Abbreviations: Mammo, mammogram; Nuc Med, nuclear medicine; US, ultrasound scan.

was judged to be obvious. For example, the erroneous text might include reference to a body part completely uninvolved in the imaging examination. Class 3 errors also involved a change in meaning as compared with the original report text, but the error was judged not to be obvious. An example would be if the phrase "enlarged lymph nodes" in the original report was transcribed as "no large lymph nodes" by the software. In general, a single error could consist of either a single word, or a multiword phrase.

### *Data Analysis*

Once the errors were classified, error rates for 3 categories of error were computed for each dictated report by dividing the total number of errors qualifying for each category by the total number of words in the report. Overall errors included all 4 error classes (class 0, 1, 2, and 3). Significant errors included only class 2 and class 3 errors. Subtle significant errors included only class 3 errors. A total of 216 individual error rates were computed (6 speakers × 12 reports/speaker × 3 error categories/report).

To observe correlation between error rates and particular reports or modalities, the averages and standard errors (SE) of the error rates were first calculated for each individual report and error category (pooling across all of the speakers). In the absence of statistically significant differences between particular reports, each report read by an individual speaker may be regarded as individual experimental measurement for that speaker (regardless of the imaging modality), and may be used to compute an average and SE characterizing the error rate for each error category for that speaker.

Next, error rates (average ± SE) for each error category for each speaker were computed (assuming no significant difference between reports, as discussed above). Error rates were then computed for groups of speakers to observe any dependence of error rate on native English speaker versus nonnative English speaker status (potentially caused by the presence of an accent), and male versus female speakers (potentially caused by differences in voice tone or pitch). A comparison also was made of the error rates of 1 of the speakers after 50-sentence and 200-sentence enrollment.

In all cases in which error rates were compared, the error rate distributions were assumed to be approximately normal, and a $t$ test was performed using a 95% confidence level.

## RESULTS

For each speaker, an evaluation was performed of the number of errors of each category observed in each dictated report. Sample results for speaker A are shown in Table 1. Next, error rates for each speaker, for each error category, and report were computed. Table 2 shows sample error rates calculated for speaker A. The error rates for each report and error category, pooled across all speakers, are shown in Table 3. No statistically significant differences between the overall error rates for the different types of reports were observed. The independence of error rate with respect to report modality is supported further by the graphic presentation of the overall error rate data in Fig 1.

Given the observed independence of error rate with respect to individual report and modality demonstrated above, the error rates for each speaker and error class pooled across modality were computed, and these results are shown in Table 4. Pooling across the entire group of 6 speakers, the error rates of overall errors, significant errors, and subtle significant errors were found to be 0.103 ± 0.033, 0.078 ± 0.034, and 0.012 ± 0.016, respectively. Also, although the 200-sentence enrollment overall and significant error rates for speaker A are seen to be somewhat smaller than the corresponding 50-sentence enrollment error rates for this

**Table 2. Error Rate Calculations for Speaker A**

| Report | Error Rates | | |
|---|---|---|---|
| | Overall | Significant | Subtle Significant |
| CT 1 | 0.063 | 0.034 | 0.010 |
| CT 2 | 0.147 | 0.132 | 0.015 |
| Mammo 1 | 0.061 | 0.030 | 0.000 |
| Mammo 2 | 0.069 | 0.034 | 0.017 |
| MRI 1 | 0.098 | 0.033 | 0.000 |
| MRI 2 | 0.087 | 0.063 | 0.000 |
| Nuc Med 1 | 0.119 | 0.075 | 0.015 |
| Nuc Med 2 | 0.077 | 0.046 | 0.000 |
| US 1 | 0.075 | 0.065 | 0.019 |
| US 2 | 0.079 | 0.048 | 0.000 |
| X-ray 1 | 0.050 | 0.000 | 0.000 |
| X-ray 2 | 0.030 | 0.000 | 0.000 |

NOTE. Similar error rate calculations also were performed for the remaining 5 speakers.

Abbreviations: Mammo, mammogram; nuc med, nuclear medicine; US, ultrasound scan.

**Table 3. Error Rates (average ± SE) Calculated for Each Report and Error Category**

| | Error Rates | | |
|---|---|---|---|
| Report | Overall | Significant | Subtle Significant |
| CT 1 | 0.078 ± 0.004 | 0.063 ± 0.004 | 0.011 ± 0.001 |
| CT 2 | 0.115 ± 0.007 | 0.100 ± 0.008 | 0.007 ± 0.001 |
| Mammo 1 | 0.101 ± 0.006 | 0.078 ± 0.008 | 0.005 ± 0.001 |
| Mammo 2 | 0.098 ± 0.005 | 0.078 ± 0.005 | 0.014 ± 0.002 |
| MRI 1 | 0.093 ± 0.005 | 0.068 ± 0.007 | 0.008 ± 0.002 |
| MRI 2 | 0.106 ± 0.005 | 0.079 ± 0.006 | 0.011 ± 0.001 |
| Nuc Med 1 | 0.119 ± 0.006 | 0.085 ± 0.007 | 0.010 ± 0.001 |
| Nuc Med 2 | 0.100 ± 0.004 | 0.063 ± 0.005 | 0.009 ± 0.001 |
| US 1 | 0.107 ± 0.005 | 0.078 ± 0.005 | 0.014 ± 0.002 |
| US 2 | 0.103 ± 0.005 | 0.079 ± 0.006 | 0.008 ± 0.002 |
| X-ray 1 | 0.125 ± 0.008 | 0.100 ± 0.009 | 0.033 ± 0.007 |
| X-ray 2 | 0.092 ± 0.004 | 0.060 ± 0.005 | 0.015 ± 0.002 |
| Average | 0.103 ± 0.013 | 0.078 ± 0.013 | 0.012 ± 0.007 |

NOTE. Error rates were pooled across all speakers. For all individual modality report error rates, n = 6. For the average error rate, n = 72.

Abbreviations: Mammo, mammogram; Nuc Med, nuclear medicine; US, ultrasound scan.

speaker, this trend was not found to be statistically significant.

Table 5 summarizes the error rates computed for each group of speakers. The native English speaker error rates are all lower than the corresponding error rates for nonnative English speakers, and these differences were found to be statistically significant for the overall and significant errors ($P$ = .009 and $P$ = .008, respectively). The error rates for the male and female speaker groups were found to exhibit no statistically significant differences.

## DISCUSSION

The overall error rate in the current study of the IBM MedSpeak continuous speech recognition system was found to be 10.3 ± 3.3%. This compares with error rates of 2.4% and 9% reported by Herman et al[9] and Zemmel et al,[11] respectively, both using the IBM VoiceType Dictation discrete speech recognition system. Teichgraber et al[15] studied the performance of the IBM ViaVoice continuous speech recognition system for dictation of CT reports and found that the recognition error rate was 3%. Arndt et al[16] studied the Philips SP6000 continuous speech recognition system and reported average error rates of 8.4% to 13.3% (which decreased to 2.4% to 10.7% after a 9-day period of initial use). Ramaswamy et al[18] used version 1.2 of the IBM Medspeak continuous

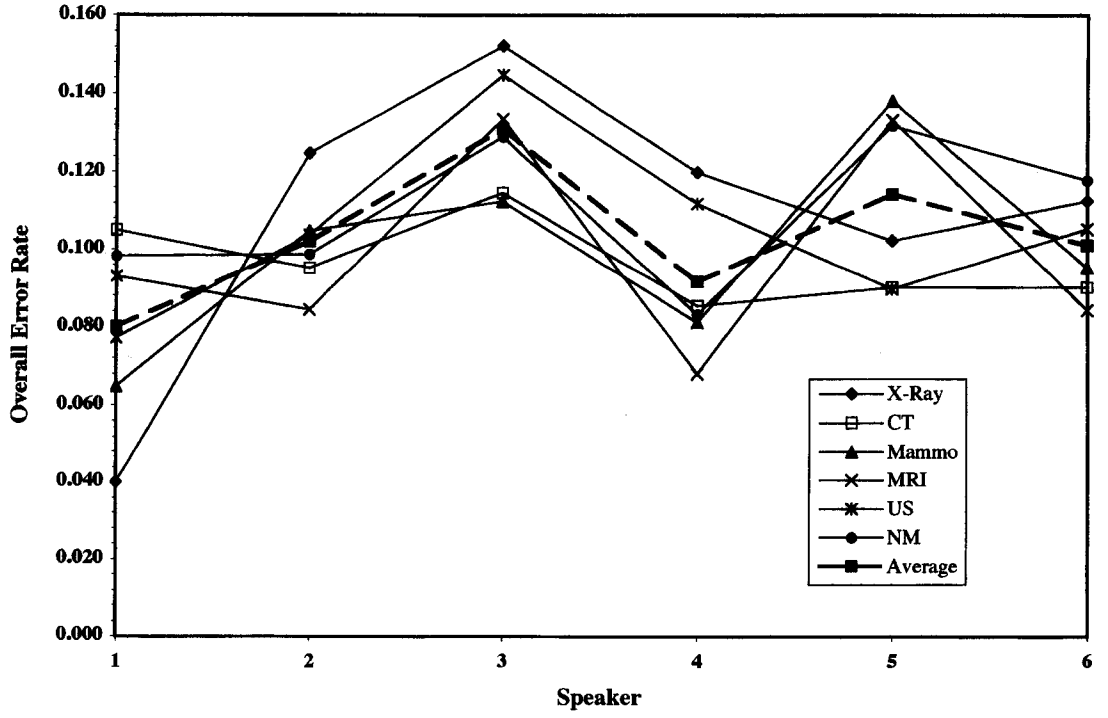


**Fig 1. Graph of overall error rates for each report modality (average of the 2 reports of each modality), plotted versus speaker. Also plotted are the average error rates for each speaker (bold line). No systematic dependence of error rate on modality is evident.**

**Table 4. Error Rates (average ± SE) Calculated for Each Speaker and Error Category**

| Category | Speaker | Error Rates | | |
|---|---|---|---|---|
| | | Overall | Significant | Subtle Significant |
| 50 sentence | A | 0.080 ± 0.031 | 0.047 ± 0.036 | 0.006 ± 0.008 |
| | B | 0.102 ± 0.021 | 0.077 ± 0.022 | 0.011 ± 0.011 |
| | C | 0.131 ± 0.031 | 0.096 ± 0.023 | 0.019 ± 0.016 |
| | D | 0.091 ± 0.030 | 0.061 ± 0.032 | 0.008 ± 0.010 |
| | E | 0.114 ± 0.032 | 0.100 ± 0.026 | 0.018 ± 0.027 |
| | F | 0.101 ± 0.030 | 0.086 ± 0.032 | 0.012 ± 0.016 |
| | Average | 0.103 ± 0.033 | 0.078 ± 0.034 | 0.012 ± 0.016 |
| 200-sentence | A | 0.071 ± 0.021 | 0.044 ± 0.019 | 0.009 ± 0.009 |

NOTE. Results are shown for the cases of 50-sentence and 200-sentence enrollment. In the 50-sentence category, n = 12 for the individual speaker error rates, and n = 72 for the average error rate. In the 200-sentence category, n = 12 for the error rates for speaker A.

speech recognition system for dictation of MR reports and concluded that the system had an average word recognition error rate of 7.3%. Our accuracy results are consistent with those of Zemmel, Arndt, and Ramaswamy, but are inferior to those of Herman and Teichgraber.

In general, discrepancies in measured accuracy between studies can result from a number of factors. Differences in error definition and accounting may contribute, especially when comparing studies of discrete and continuous speech recognition systems. For the discrete speech case, each word in the report potentially may contribute to only a single error in the transcribed report, whereas in the continuous speech case, because phrases involving several words may be incorrectly recognized, each word can contribute to errors in a wide variety of ways. Also in some study designs, discrepancies arising from words that were mispronounced or slurred together may not be counted as errors (eg, in Herman[9]), whereas in the current study, all observed discrepancies between the original and transcribed report were counted as errors. It is conceivable that studies that are focused on a particular modality area (eg, MR or CT) may "play to the strength" of a particular speech recognition product either inadvertently or by design, and produce superior recognition results, whereas a "multimodality" study of the same product could find lower accuracy. Differences in the particular speech recognition products and versions evaluated also may tend to produce variable accuracy findings. Finally, it has been documented that some speech recognition systems are able to improve their accuracy over time,[16] so the amount of experience with the system by the evaluators before the study also may affect the measured accuracy.

**Table 5. Error Rates (average ± SE) Calculated for Groups of Speakers for Comparison of Native English Versus Nonnative English Speakers, and Male Versus Female Speakers**

| Speaker Groups | Error Rates | | | No. |
|---|---|---|---|---|
| | Overall | Significant | Subtle Significant | |
| Native | 0.097 ± 0.031 | 0.071 ± 0.035 | 0.011 ± 0.016 | 48 |
| Nonnative | 0.116 ± 0.034 | 0.091 ± 0.028 | 0.015 ± 0.016 | 24 |
| Men | 0.104 ± 0.035 | 0.073 ± 0.034 | 0.012 ± 0.013 | 36 |
| Women | 0.102 ± 0.031 | 0.082 ± 0.034 | 0.012 ± 0.019 | 36 |

In certain applications, the rate of significant errors may be more relevant than the overall error rate, because these are the errors that would impose patient management consequences. The rate of significant errors was found to be 7.8 ± 3.4% (still approximately 3 times greater than the overall error rate previously reported for the discrete speech system). On average, the rate of incidence of significant errors in an 87-word report (the observed mean word length of our test set of reports) would be about 7. Each of these errors would require manual correction by the system user. Of course, this number will be lower in practice areas that tend to produce more concise reports.

The rate of subtle significant errors (those errors in which detection would require an especially careful proofreading of the transcribed report) was computed to be 1.2 ± 1.6%. As evidenced by the very large SE value relative to the average, few errors in this category were observed. On average, the rate of incidence of subtle significant errors in an 87-word report would be about 1. Of the 84 total reports transcribed by the system in this study, 50

had one or more subtle significant errors. The efficiency of the proofreading process necessary for error detection would be much improved if an indication of the confidence of the speech recognition algorithm in its interpretation of the audio could be provided by the program. This could be indicated using color-coded highlighting of the report text. The radiologist could rapidly skim areas of high confidence, and focus mainly on areas of low confidence. Such cues commonly are provided by human transcriptionists (eg, using question marks to delimit relevant areas of text).

Statistically significant differences in the accuracy were observed for the overall and significant errors as a function of the native English speaker status, although the error rates were fairly similar to one another. No statistically significant differences were seen between the male and female speaker groups. Zemmel et al[11] also studied the differences between male and female speakers for a discrete speech recognition system and found no statistically significant differences. Also, no statistically significant difference in the accuracy of speaker A was observed with respect to the 50-word or 200-word enrollment procedures.

The potential for use of any speech recognition technology will be a function of the overall transcription accuracy, as well as a number of other significant factors.[14,21-26] These additional considerations include the convenience of the existing transcription operation to the radiologists, implementation issues, the convenience of the computer-based transcription system (including the entry of information identifying the specific patient and examination, accuracy with respect to capturing ancillary information, and the efficiency of error detection and correction), the software, hardware and maintenance of the speech recognition system, the reading room environment (very noisy or relatively quiet), user experience with the system, the current examination turnaround time to the clinical physicians, legal and financial implications, and the presence of other electronic systems such as RIS and PACS.

As an example of the importance of these considerations, for the majority of work done in the Department of Diagnostic Radiology at Mayo Clinic (Rochester, MN), the convenience of the current transcription systems is quite high (involving either direct dictation to a human transcriptionist, or voice recording with subsequent playback and transcription to text by a human transcriptionist). There also is a generally high level of satisfaction with the current manual system of delivery of hardcopy examinations (including final report and images) to the clinical floors within an average of about 2 to 3 hours after examinations completion. With the configuration of the MedSpeak/Radiology system we studied, we encountered several factors, which made it inconvenient to use the system routinely. Although the current study did not formally include the dictation of numeric indication and diagnosis codes (important to allow searching of a radiology report database), anecdotal reports indicate that the system was not able to capture this information as accurately as it could the main report text. In practice, these codes would need to be entered manually. Similarly, it was found to be somewhat awkward to enter the necessary patient identification information and to move the report text into the RIS. These problems could be addressed through the development of an interface between the transcription software and the RIS, but given the current manual system of hardcopy delivery of examinations to the clinical floors, the interface would have only a marginal impact on examination turnaround time. As the use of PACS in the practice increases, and the system for electronic delivery of radiology examination information to clinical floors is more widely available, the potential examination turnaround time will decrease substantially from current levels, and the impact of an automated speech recognition system on turnaround time will increase greatly. At that point, the practice will be much more likely to consider the routine use of an automated speech recognition system, even if the transcription efficiency of the radiologist would be somewhat lowered. This is because of the very large potential improvement to the overall delivery of patient care. The current level of overall accuracy of the MedSpeak/Radiology system probably is sufficient for this type of scenario to proceed.

One area in the Mayo Clinic (Rochester) practice that is a strong candidate for immediate application of the speech recognition system is after-hours coverage of the emergency room. The radiologist must type the report manually into the RIS in this situation. The practical drawbacks associated with use of the speech recognition system noted above are overshadowed by the inconveniences of the current practice. Initial use of the

MedSpeak/Radiology system for this application has resulted in quicker generation and clinical distribution of the report.

The study design utilized in this work required a reasonably low investment of time on the part of 6 subjects (here radiologists and physicists). The time invested by each participant was less than about 35 minutes including the time necessary to complete the enrollment process. In spite of this, the study considered multiple imaging modalities, native English speaker status, gender, as well as different error categories, and also was able to show relatively small differences in error rate (approximately 2%) as statistically significant. This study design should be useful in efficiently comparing different speech recognition products based on error rate and documenting changes in error rates induced by extended registration processes or ongoing system "learning" (if these features are present in the system of interest). Our experimental design in this work used new users for the dictation of the reports. However, this is not a requirement for testing speech recognition programs in general. Experienced users could also use the same methodology described here to compare different speech recognition products or to evaluate other variables such as the impact of hardware or software upgrades, the efficacy of additional training, and the efficacy of optimizing recognition software settings. Some speech recognition systems are able to improve their accuracy over time[16] and with more experienced users. Our users were chosen randomly to reasonably represent our clinical practice. We are not suggesting that this methodology be used to compare speech recognition systems between different institutions but rather it be used within an institution to compare accuracy and efficiency of any speech recognition system, initially and over a period of time. When comparing speech recognition systems between institutions, other variable factors such as reading room environment, user variability, software and hardware systems, and general characteristics of the reports themselves (short, verbose), as mentioned previously, would have to be considered, which would be difficult.

## CONCLUSIONS

The MedSpeak/Radiology continuous speech recognition software was installed and run on a modest personal computer platform. It was evaluated by 6 speakers using a test set of twelve radiology reports. Our findings indicate that the software is approximately 90% accurate when all types of errors are considered. The recognition accuracy for the software tested in this paper displayed no dependence on the type of report or the sex of the speaker. Statistically significant differences in recognition accuracy were seen when groups of native and nonnative English speakers were compared, although the accuracy values were similar. Enrolling with the maximum number of sentences had no statistically significant effect on recognition accuracy when compared with results obtained using the minimum enrollment procedure. Practical implementation issues (rather than recognition accuracy) currently limit the widespread routine use of the system in radiology, although niche applications (eg, after-hours emergency room interpretation) likely will benefit from use of the system. It is expected that the use of continuous speech recognition systems interfaced with an RIS, and used along with PACS will remove the major practical impediments to routine applications. Our methodology provides a convenient and efficient way to compare the accuracy of different speech recognition products, changes in accuracy over time, and impact of other system factors on accuracy.

## ACKNOWLEDGMENT

The authors acknowledge the assistance of William S. Harmsen (Section of Biostatistics, Mayo Clinic, Rochester) with the statistical design and analysis for this study.

## REFERENCES

1. Leeming BW, Porter D, Jackson JD, et al: Computerized radiologic reporting with voice data-entry. Radiology 138:585-588, 1981

2. Robbins AH, Horowitz DM, Srinivasan MK, et al: Speech-controlled generation of radiology reports. Radiology 164:569-573, 1987

3. Robbins AH, Vincent ME, Shaffer K, et al: Radiology reports: Assessment of a 5,000-word speech recognizer. Radiology 167:853-855, 1988

4. Smith NT, Brien RA, Pettus DC, et al: Recognition accuracy with a voice-recognition system designed for anesthesia record keeping. J Clin Monit 6:299-306, 1990

5. Massey BT, Geenen JE, Hogan WJ: Evaluation of a voice recognition system for generation of therapeutic ERCP reports. Gastrointest Endosc 3:617-620, 1991

6. Holbrook JA: Generating medical documentation through voice input: the emergency room. Top Health Rec Manage 12:49-57, 1992

7. Reed RA: Voice recognition for the radiology market. Top Health Rec Manage 12:58-63, 1992

8. Clark S: Implementation of voice recognition technology at Provenant Health Partners. J AHIMA 65:34-38, 1994

9. Herman SJ: Accuracy of a voice-to-text personal dictation system in the generation of radiology reports. AJR 165:177-180, 1995

10. Meijer GA, Baak JP, Van Diest PJ, et al: Text processing by digital voice recognition. Anal Quant Cytol Histol 18:261-266, 1996

11. Zemmel NJ, Park SM, Schweitzer J, et al: Status of voicetype dictation for windows for the emergency physician. J Emerg Med 14:511-515, 1996

12. Rosenthal DI, Chew FS, Dupuy DE, et al: Computer-based speech recognition as a replacement for medical transcription. AJR 170:23-25, 1998

13. Korn K: Voice recognition software for clinical use. J Am Acad Nurse Pract 10:515-517, 1998

14. Mehta A, Dreyer KJ, Schweitzer A, et al: Voice recognition-an emerging necessity within radiology: Experiences of the Massachusetts General Hospital. J Digit Imaging 11:20-23, 1998 (suppl 2)

15. Teichgraber UK, Ehrenstein T, Lemke M, et al: Automated speech recognition for the generation of medical records in computed tomography. [Article in German, Abstract in English]. Rofo. Fortschritte auf dem Gebiete der Rontgenstrahlen und der Neuen Bildgebenden Verfahren 171:369-399, 1999

16. Arndt H, Petersein J, Stockheim D, et al: Automated speech recognition in diagnostic radiology. [Article in German, Abstract in English]. Rofo. Fortschritte auf dem Gebiete der Rontgenstrahlen und der Neuen Bildgebenden Verfahren 171: 400-404, 1999

17. Zafar A, Overhage JM, McDonald CJ: Continuous speech recognition for clinicians. J Am Med Inform Assoc 6:195-204, 1999

18. Ramaswamy MR, Chaljub G, Esch O, et al: Continuous speech recognition in MR imaging reporting: Advantages, disadvantages and impact. AJR 174:617-622, 2000

19. Schwartz LH, Kijewski P, Hertogen H, et al: Voice recognition in radiology reporting. AJR 169:27-29, 1997

20. Bergeron BP: Usable voice-recognition technology. It's finally arrived! Postgrad Med 102:39-44, 1997

21. Bergeron BP: Voice recognition: An enabling technology for modern health care? Proceedings/AMIA Annual Fall Symposium, Philadelphia, PA, Hanley & Belfus, 1996, pp 802-806.

22. Seltzer SE, Kelly P, Adams DF, et al: Expediting the turnaround of radiology reports in a teaching hospital setting. AJR 168:889-893, 1997

23. Kovesi T: Dictation software for MDs improving but frustration still part of the program. Can Med Assoc J 158:1059-1060, 1998

24. Threet E, Fargues MP: Economic evaluation of voice recognition for the clinicians' desktop at the Naval Hospital Roosevelt Roads. Mil Med 164:119-126, 1999

25. Pavlicek W, Muhm JR, Collins JM, et al: Quality-of-service improvements from coupling a digital chest unit with integrated speech recognition, information, and picture archiving and communication systems. J Digit Imaging 12:191-197, 1999

26. Houston DJ, Rupp FW: Experience with implementation of a radiology speech recognition system. J Digit Imaging 13:124-128, 2000