The Long-Term Adoption of Speech Recognition in Medical Applications

Michael A. Grasso
George Washington University School of Medicine
grasso@gwu.edu

Abstract

This paper presents a survey on the long-term adoption of speech recognition in medical applications. Thirty-one participants who authored papers on medical speech recognition applications responded to the survey. The participants viewed speech technology more favorably today than when they originally published their papers. However, the adoption of speech applications did not always correspond with their enthusiasm. The survey suggested that hands-busy, eyes-busy, and mobility requirements are not always enough to offset current limitations in speech technology. There may need to be other benefits, such as decreased medical costs and increased quality of care, or other factors, such as using a limited vocabulary.

1. Introduction

The development of a conversational computer has been an elusive goal for more than 30 years. Despite considerable advances in computer technology, the keyboard and mouse are still the principal means of entering data. While improvements have taken place, speech technology has a number of limitations that keep it out of the mainstream.

It is widely believed that speech recognition technology works best when there is a compelling reason to use it [1]. Examples of this include hands-busy, eyes-busy, and mobility-required applications [2]. While these observations are intuitive, little work has been done to empirically study the limits of these boundaries. To better understand the tenor of speech recognition technology in the medical field, and provide empirically based insight on the best ways to apply it, this paper reports on the disposition and implementation of speech-driven medical applications over the last ten years.

2. Background

Speech recognition systems provide computers with the ability to identify spoken words and phrases. As shown in Figure 1, a computer receives an analog signal through a microphone. The analog signal is converted to a digital waveform. The digital waveform is compared to a database of known waveforms for all phonemes. A phoneme is the smallest unit of speech and represents an individual sound. Finally, sequences of phonemes are assembled into words and phrases using a stochastically based lexicon.

Speech interfaces have a number of unique characteristics when compared to traditional modalities. The most significant is that speech is temporary. Once a phrase is spoken, auditory information is no longer available. This places extra memory burdens on the user and severely limits the ability to scan, review, and cross-reference information. Speech can be used at a distance, which makes it ideal for hands-busy and eyes-busy situations. It is omnidirectional and can communicate with multiple users, which has privacy implications. There are also problems related to anthropomorphism, where users tend to overestimate the capabilities of a speech interface and are tempted to treat the device as another person [3].

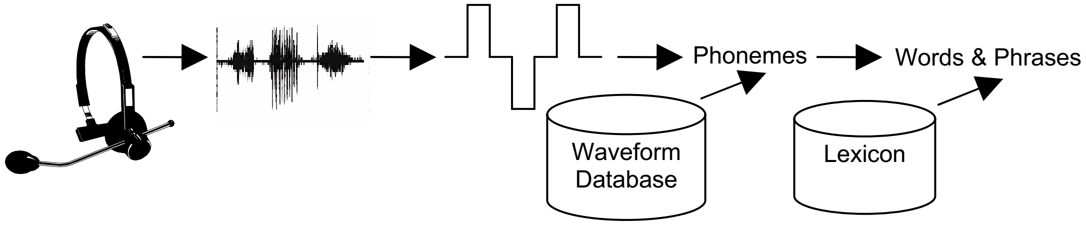

Figure 1. Speech Recognition Algorithm

At the same time, speech recognition systems often carry technical limitations, such as speaker dependence, continuity, and vocabulary size. Speaker-dependent systems must be trained by each individual user, but typically have higher accuracy rates than speaker-independent systems, which can recognize speech from any person. Continuous-speech systems recognize words spoken in a natural rhythm, while isolated-word systems require a deliberate pause between each word. Although more desirable, continuous-speech is harder to process, because of the difficulty in detecting word boundaries. Vocabulary size can vary anywhere from 20 words to more than 40,000 words. Large vocabularies cause difficulties in maintaining recognition accuracy, but small vocabularies can impose unwanted restrictions. A more thorough review can be found elsewhere [4].

Recent trends in speech recognition systems have been geared toward large vocabulary, speaker independence, and continuous recognition. In the medical field, these innovations have primarily been incorporated into dictation systems for the development of reports in areas like radiology, pathology, and endoscopy [5,6,7,8,9,10]. With this approach, a physician can dictate clinical narratives directly into a computer, eliminating the need for a transcription service and thus decreasing turnaround time.

The use of free-form dictation with a speech recognition engine may be sufficient for report generation. However, the ambiguous nature of clinical narratives makes it difficult to quantify and analyze this information [11]. Structured speech input can address these limitations by recognizing speech as a series of coded and quantifiable pieces of information [12,13].

Other medical applications of speech interfaces include software command and control. In this context, speech is often combined with other input devices to form a multimodal interface [14,15]. The advantage of this approach is that the reciprocal strengths of one input modality can offset the weaknesses of another [16]. A related application is to use speech input for instrument control, such as during surgical procedures [17] or for the physically impaired [18].

3. Materials and Methods

A survey was distributed by email to people who authored papers on speech-driven medical applications between 1992 and 2002. The list of journals and conference publications was limited to those that were indexed by the National Library of Medicine, and included several peer-reviewed medical and medical informatics publications, including all peer-reviewed publications of the American Medical Informatics Association. The authors were asked by email to fill out an online survey that focused on trends in speech recognition technology in the medical field. The goal of the survey was to compare people's views of speech technology to the way they actually implemented speech-driven systems.

Approximately 40 percent of the primary authors who were contacted responded to the survey (31 of 78). Ten of the participants had medical degrees (MD, DO), 14 had doctoral degrees (PhD, PharmD, DSc), and 7 had Bachelors or Masters degrees. Sixteen of the participants were from academic institutions, 8 were from industry, and 7 were from medical institutions. The average timeframe between when their articles were published and when they answered the survey was 4.7 years, with a range of 1 to 10 years.

Participants were asked to rate their view of speech technology as being accurate, dependable, efficient, mature, and useful. They answered these questions based on their initial views at the time their papers were published, and based on their current views. Each question was answered on a scale from 1 to 9, where a 1 represented strong disagreement, a 5 was neutral, and a 9 was strong agreement.

Participants also categorized the way speech recognition technology was being used at their organizations in four key areas: dictation and report generation, structured data entry, software command and control, and instrument control. Each area was rated on a scale of 1 to 9, where a 1 represented no usage, a 5 was moderate usage, and a 9 was significant usage. They answered these questions based on how the technology was initially used (at the time of their publication), the way it is currently used, and they way it is predicted to be used by their organizations in the future.

4. Results

The mean response for each question across all participants was computed and normalized on a scale of 1 to 9. A higher value was indicative of higher acceptance or use. The overall trend was that acceptance of speech recognition technology increased from an initial view of 4.95 to a current view of 5.95. At the same time, the adoption of speech recognition technology decreased slightly from 2.62 to 2.54. When looking to the future, usage was predicted to increase to 3.43.

A breakdown of user acceptance by question is shown in Figure 2. For each question, acceptance increased from the initial period to the present. All values were statistically significant using a two-tailed paired t-test ($p < 0.05$).

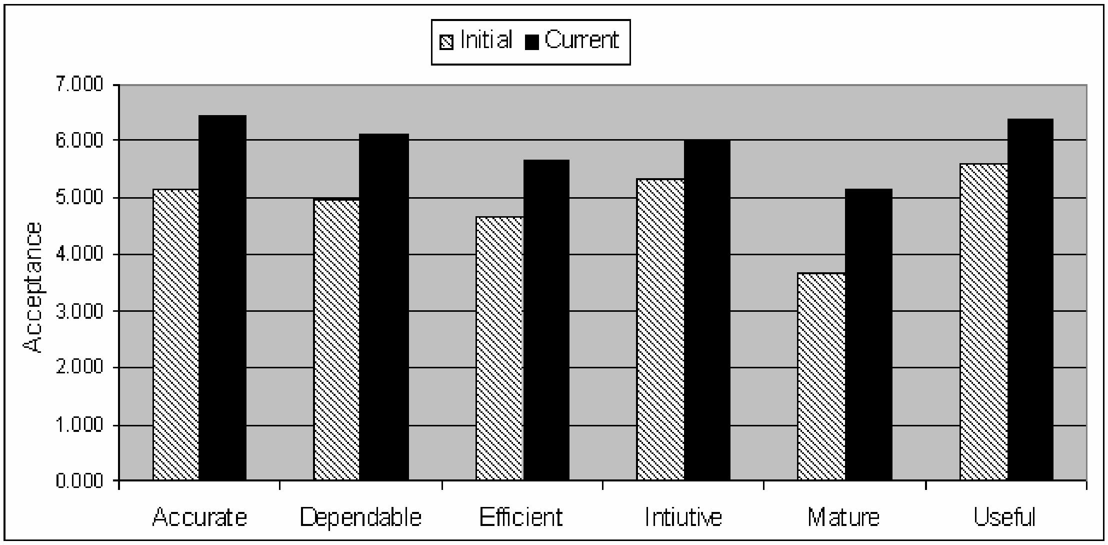

Figure 2. Acceptance by Question

An acceptance score of 5 was considered neutral. The only questions answered positively for the initial period were that speech was intuitive (mean = 5.53) and useful (mean = 5.65). The initial questions on speech being accurate (mean = 5.00), dependable (mean = 4.94), efficient (mean = 4.82), and mature (mean = 3.77) were answered as neutral or negative. For the current period, all questions were answered positively: accurate (mean = 6.35), dependable (mean = 6.06), efficient (mean = 5.71), intuitive (mean = 6.06), mature (mean = 5.18), and useful (mean = 6.35).

Figure 3 contains a breakdown of how speech recognition technology is being used in medical applications at each participant's organization. The adoption of speech technology increased for dictation and for instrument control from the initial period to the current period, but decreased for structured data entry and for software control. Participants predicted that the adoption of speech technology would increase in all areas in the future. All values were statistically significant using a two-tailed paired t-test (p < 0.05), except initial usage to present usage.

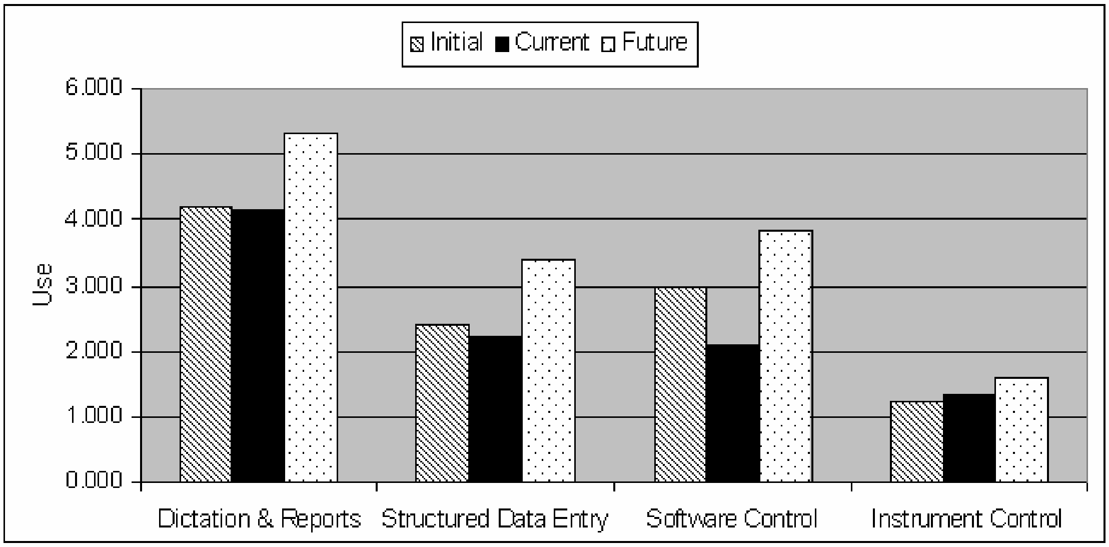

Figure 3. Adoption of Speech Recognition

A score of 5 was considered moderate usage. For the initial period, no question was answered above the moderate level: dictation (mean = 3.94), structured data entry (mean = 2.47), software control (mean = 2.82), and instrument control (mean = 1.24). For the current period, no question was answered above the moderate level: dictation (mean = 4.35), structured data entry (mean = 2.29), software control (mean = 2.18), and instrument control (mean = 1.35). In the future, dictation was predicted to be above the moderate level (mean = 5.12), while structured data entry (mean = 3.24), software control (mean = 3.71), and instrument control (mean = 1.65) were not.

The results were analyzed for linear relationships using correlation coefficients (r), with the statistical significance verified by two-tailed paired t-tests. Overall acceptance did not correlate with use of speech technology (r = 0.14, p > 0.05). However, the participants' current acceptance had a positive linear correlation with both current use (r = 0.49, p < 0.05) and predicted use (r = 0.62, p < 0.01). The significance of current acceptance to overall use was borderline (r = 0.42, p < 0.1).

5. Discussion

The survey revealed that optimism exists on the applicability of speech recognition technology for medical applications. The technology is viewed more favorably today than when the participants wrote their articles. All participants predicted that their organizations will increase their use of the technology in the future. At the same time, their actual adoption of speech applications did not always concur with their optimism. The current usage was more variable, and overall it decreased slightly from the time the participants' papers were published. This may explain why the t-test result was not significant when comparing the initial use to the current use of speech technology.

One notable exception was the increased use of speech recognition for dictating clinical reports. This process tends to have recognition accuracy rates between 80% and 95%, which can increase the time needed to dictate reports by 25% or more. This would normally make the adoption of speech technology less viable [19]. However, this approach is considerably less expensive than using a transcription service, and can cut the turnaround time from days to a matter of hours [20]. These added benefits, which can reduce medical costs and increase the quality of care, seem to make speech recognition a viable option, in spite of accuracy problems.

The adoption of this technology also increased when using a speech interface to control a medical instrument. An example of this is a surgeon who controls a camera with speech commands while operating on a patient. Note that a system of this type might use no more than 10 commands to control a camera. This concurs with the general view that speech-driven applications are most suitable in hands-busy environments with limited vocabularies [21].

The survey showed that the adoption of speech technology decreased for structured data entry. In contrast to free-form dictation, these applications process speech input as coded and quantifiable pieces of information [12]. This approach has been shown to increase recognition accuracy over free-form dictation [14,22]. Most applications developed in this category tend to be research oriented, are not used as an alternative to transcription services, and are not under the same time constraints as clinical reports. This suggests that the advantage of hands-busy data entry by itself is not always sufficient to justify the use of speech technology, even when the approach can decrease speech recognition errors. There may need to be other incentives to justify the current limitations of the technology.

The use of speech technology for software command and control also decreased. An example of this is using speech input to control your word processor or other desktop computer applications. This is probably the most oversold area of speech technology. Most every speech recognition environment comes with command and control capabilities. However, simply adding speech to an existing user interface can decrease system integrity or create integration discontinuity [23]. It is normally best to design speech-driven interfaces from scratch, to examine user interaction from this new perspective.

The response rate of 40 percent creates the possibility of a response bias. In addition, survey responses were limited to primary authors of papers on medical speech applications. It is also important to note that the data are dependent on personal reporting and are subject to recall bias.

6. Conclusion

This paper presented the results of a survey on the disposition and implementation of speech-driven medical applications. It correlated people’s views of speech technology to the way they are actually implementing it. Thirty-one participants who authored papers on medical speech recognition applications responded. The participants viewed speech technology more favorably today than when they published their papers. However, the adoption of speech applications did not always correspond with their enthusiasm. The survey also suggested that hands-busy, eyes-busy, and mobility requirements are not always enough to offset the current limitations in speech technology. There may need to be other benefits, such as decreased medical costs and increased quality of care, or other factors, such as using a limited vocabulary.

7. Acknowledgements

Michael A. Grasso (http://home.gwu.edu/~grasso) is a medical student at the George Washington University School of Medicine in Washington, DC. He also holds a PhD in computer science from the University of Maryland Baltimore County.

This research was supported by grant 2R44RR07978 from the National Center for Research Resources at NIH.

8. References

1. Lai J. Conversational Interfaces. Communications of the ACM 2000 Sep;43(9):24-27.
2. Shneiderman B. The Limits of Speech Recognition. Communications of the ACM 2000 Sep;43(9):63-65.
3. Jones, D. M., Hapeshi, K. and Frankish, C. Design Guidelines for Speech Recognition Interfaces. Applied Ergonomics 1990;20:40-52.
4. Peacocke, R. D. and Graf, D. H. An Introduction to Speech and Speaker Recognition. IEEE Computer 1990;23(8):26-33.
5. Callaway EC, Sweet CF, Siegel E, Reiser JM, Beall DP. Speech Recognition Interface to a Hospital Information System Using a Self-Designed Visual Basic Program: Initial Experience. Journal of Digital Imaging 2002 Mar;15(1):43-53.
6. Hollbrook JA. Generating Medical Documentation Through Voice Input: The Emergency Room. Topics in Health Records Management 1992;12(3):58-63.
7. Lai J, Vergo J. MedSpeak: Report Creation with Continuous Speech Recognition. Proceedings of the Conference on Human Factors in Computing Systems (CHI'97), 1997;:431- 438.
8. Klatt EC. Voice-Activated Dictation for Autopsy Pathology. Computers in Biology and Medicine 1991:21(6):429-433.
9. Massey BT, Geenen JE, Hogan WJ. Evaluation of a Voice Recognition System for Generation of Therapeutic ERCP Reports. Gastrointestinal Endoscopy 1991;37(6):617-620.
10. Korn K. Voice Recognition Software for Clinical Use. Journal of the American Academy of Nurse Practitioners 1998;10(11):515-517.
11. Tange HJ, Hasman A, De Vries Robbe PF, Schouten HC. Medical Narratives in Electronic Medical Records. International Journal of Medical Informatics 1997;46:7-29.
12. Grasso MA. Structured Speech Input for Clinical Data Collection. 15th IEEE Symposium on Computer Based Medical Systems, (CBMS 2002), 2002;:199-204.
13. Teel MM, Sokolowski R, Rosenthal D, Belge M. Voice-Enabled Structured Medical Reporting. Proceedings of the Conference on Human Factors in Computing Systems (CHI'98), 1998;:595-601.
14. Grasso MA, Ebert DS, Finin TW. The Integrality of Speech in Multimodal Interfaces. ACM Transactions on Computer-Human Interaction 1998;5(4):303-325.
15. McMillan PJ, Harris JG. Datavoice: A Microcomputer-Based General Purpose Voice-Controlled Data-Collection System. Computers in Biology and Medicine 1990; 20(6):415-419.
16. Cohen PR. The Role of Natural Language in a Multimodal Interface. In Proceedings of the ACM Symposium on User Interface Software and Technology (UIST 1992), 1992;:143-149.
17. Rossi L, Sacerdoti D, Billi B, Lesnoni G, Orciuolo M, Rossi T, Sacerdoti D, Bertollini L. Automatic Speech Recognition in Vitreo-Retinal Surgery. European Journal of Ophthalmology 1996 Oct-Dec;6(4):454-459.
18. Lin CL, Won RM, Luh JJ, Lee MH, Kuo TS, Ru CT. A Radio Controller Using Speech for the Blind. Critical Reviews in Biomedical Engineering 2000;28(3-4):429-433.
19. Karat CM, Halverson C, Horn D, Karat J. Patterns of Entry and Correction in Large Vocabulary Continuous Speech Systems. Proceedings of the Conference on Human Factors in Computing Systems (CHI'99), 1999;:568-575.
20. Anonymous. Speech Recognition Systems. Are They up to the Task? Health Devices 2002 Feb;31(2):65-71.
21. Grasso MA. Automated Speech Recognition in Medical Applications. M.D. Computing 1995;:12(1):16-23.
22. Oviatt SL. Multimodal Interfaces for Dynamic Interactive Maps. Proceedings of the Conference on Human Factors in Computing Systems (CHI'96), 1996;:95-102.
23. Wulfman CE, Isaacs EA, Webber BL, Fagan LM. Integration Discontinuity: Interface Users and Systems. Tech. Report KSL-88-12, Knowledge Systems Laboratory, Stanford University, Palo Alto, California, 1988.