Towards spoken clinical-question answering: evaluating and adapting automatic speech-recognition systems for spoken clinical questions

Feifan Liu,[1] Gokhan Tur,[2] Dilek Hakkani-Tür,[3] Hong Yu[1,4]

[1]Department of Health Sciences, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA
[2]Speech Technology & Research Laboratory, Information and Computing Sciences Division, SRI International, Menlo Park, California, USA
[3]Speech Group, International Computer Science Institute, Berkeley, California, USA
[4]Department of Computer Science, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA

Correspondence to
Dr Hong Yu, 2400 E Hartford Ave, Room 939, Milwaukee, WI 53211, USA;
hongyu@uwm.edu

Received 18 March 2010
Accepted 4 May 2011
Published Online First
24 June 2011

ABSTRACT

Objective To evaluate existing automatic speech-recognition (ASR) systems to measure their performance in interpreting spoken clinical questions and to adapt one ASR system to improve its performance on this task.

Design and measurements The authors evaluated two well-known ASR systems on spoken clinical questions: Nuance Dragon (both generic and medical versions: Nuance Gen and Nuance Med) and the SRI Decipher (the generic version SRI Gen). The authors also explored language model adaptation using more than 4000 clinical questions to improve the SRI system's performance, and profile training to improve the performance of the Nuance Med system. The authors reported the results with the NIST standard word error rate (WER) and further analyzed error patterns at the semantic level.

Results Nuance Gen and Med systems resulted in a WER of 68.1% and 67.4% respectively. The SRI Gen system performed better, attaining a WER of 41.5%. After domain adaptation with a language model, the performance of the SRI system improved 36% to a final WER of 26.7%.

Conclusion Without modification, two well-known ASR systems do not perform well in interpreting spoken clinical questions. With a simple domain adaptation, one of the ASR systems improved significantly on the clinical question task, indicating the importance of developing domain/genre-specific ASR systems.

INTRODUCTION

Studies have shown that clinicians have many questions when seeing patients.[1–7] Table 1 shows a subset of questions posed by clinicians. Identifying possible answers to such questions will support the practice of evidence-based medicine[8] and, as a result, may improve the quality of patient care.[9–11] With that goal in mind, we are developing a clinical question answering (QA) system called AskHERMES—Help clinicians Extract and aRticulate Multimedia information from literature to answer their ad hoc clinical quEstionS—which automatically extracts information needs from ad hoc clinical questions, returns relevant documents, extracts relevant answers, and summarizes and formulates answers in response to these questions. AskHERMES has the potential to help clinicians effectively identify answers they need at the point of patient care.

One challenge that clinical questions pose for an automatic QA system is that they are typically long and complex. Table 1 presents examples randomly selected from the 4654 questions in the ClinicalQuestion data, which were collected from healthcare providers across the USA.[1 12–14] The average number of word tokens for each question in the collection is 20, and busy clinicians rarely have the time to type questions of such length into computers or portable devices, such as personal digital assistants, as question-answering systems have traditionally required.

Speech is a fundamental (and perhaps the most important and natural) modality of interaction, providing an efficient way to address the aforementioned challenge in QA systems. For a clinician, a speech interface to QA would save time and also allow for more natural and easy interaction during searches for answers to questions. A speech interface would also support QA via cell phone and other portable devices in cases in which there is no computer access; such situations include combat zones and ambulance delivery. Moreover, a speech interface could circumvent potential confusion resulting from spelling errors, alternative spellings, and abbreviations that often accompany the use of long and complex medical terms in text-based QA. This study reports our evaluation of one state-of-the-art automatic speech-recognition (ASR) tool and a heavily used off-the-shelf commercial ASR system on spoken clinical questions, the subsequent domain-specific adaptation of one of these systems, and the evaluation of the adapted system.

RELATED WORK

Most ASR work in the clinical domain focuses on medical dictation. Such work can be generally grouped into two categories: performance evaluation of multiple speech-recognition software products [15–17] and usability studies.[18–21] Zafar *et al*[15] reported training times and accuracy rates on different ASR systems when default and additional medical dictionaries were used. They also reported that ambient noise (within reason) had no real effect on the recognition accuracy. In their later work,[16] they identified nine categories of errors committed by Nuance Dragon (4.0) on clinical notes. Similarly, Devine *et al*[17] compared the out-of-box performance of three commercially available continuous speech-recognition software packages for dictating medical progress notes and discharge summaries. They found that IBM ViaVoice 98 with General Medicine Vocabulary had the lowest mean error rate (7.0–9.1%), while Nuance Dragon Medical (version 3.0) had the highest (14.1–15.2%). In their studies,[15–17] the text was read by the same speaker to different speech recognizers at different time, but the

Table 1 Subset of clinical questions collected by Ely and associates[1]

Question type	Sample questions
'What …' (48%)	1. What should you do for asymptomatic carotid bruits or bruits in general? Folic acid? Vitamin b_1? Aspirin?
'How …' (15%)	3. How long should you leave a patient on coumadin and heparin? Can I stop the heparin as soon as the protime is therapeutic?
'Do …' (7%)	4. Does this patient with a 3-week-old fracture of the distal head of the fifth metatarsal need any treatment?
'Can …' (4%)	5. Can Lorabid cause headaches?
Others (48%)	6. This woman takes Premarin 0.625 mg for osteoporosis prophylaxis but she got sore breasts. Is there a smaller size premarin like 0.3 mg? 7. Colon cancer screening. I had sent one of my 43-year-old patients to gastroenterology for screening colonoscopy because her father had colon cancer in his 60s. They sent her back, saying we should start screening at age 50?

The left column represents generic question proportions. For example, 'What,' 'How,' 'Do,' and 'Can' account for 2231 (or 48%), 697 (or 15%), 320 (or 7%), and 187 (or 4%), respectively, of all clinical questions. Question examples (1–6) are in the right column. We have kept the questions in their original form, preserving misspellings and other types of errors. Many questions (eg, Question 7) are in a very informal, conversational style.

speaker's pronunciation is likely to change over the time even for the same words. Therefore, the acoustic properties of the test data were expected to be different for different speech recognizers.

Some published studies[18 19 21] have presented the results of ASR software being used for transcribing part of the work, with the rest being transcribed via humans. One study[18] reported that computerized speech recognition may be an acceptable alternative to human medical transcription for producing outpatient notes, while other studies found the additional cost incurred by using an automatic speech recognizer unacceptable in their clinical documentation process.[19 21]

We found limited ASR work in the clinical domain on spontaneous speech, which is the type of speech that is the focus of this study. ASR systems for spontaneous speech have been developed in other domains, however. One study in another domain[22] showed that spontaneous speech effects significantly degraded recognition performance. A multiword model was developed for modeling repetitions for recognition of conversational telephone speech,[23] leading to an absolute WER reduction of 2.0%, from 42.1% to 40.1%, on already well-trained acoustic and language models. Using unsupervised language model adaptation, Niesler and Willett[24] reported the WER of 35.7% on their lecture speech test set, Tur and Stolcke[25] reported the WER of 12.1% on meeting speech data, and Liu *et al*[26] achieved a character error rate of 14.2% on Chinese Mandarin speech data.

Since the commercially available Nuance Dragon ASR systems were introduced in 1992,[27] they have been widely used in hospitals for physician dictation, especially in the radiology domain. Surprisingly, only a few studies have evaluated the performance of Nuance Dragon ASR systems. One study showed that Nuance Dragon was successful in interpreting dictated cardiological reports.[28] Other studies have shown that Nuance Dragon ASRs are not user-friendly and show disappointing performance in some clinical subdomains. Havstam *et al*,[20] for example, concluded that it is time-consuming to learn to use Nuance Dragon, and Issenman and Jaffer[29] concluded that Nuance Dragon (6.0) was disappointing because, despite its steep learning curve, its recognition performance was poor, effectively limiting its broad acceptance among physicians.

Previous ASR evaluations in the clinical domain all measured performance on continuous speech that was rehearsed or read aloud. Our goal was to evaluate ASR systems on spontaneously spoken questions, inspired by a recent study[30] in an inpatient setting demonstrating the feasibility of voice capturing medical residents' clinical questions in spontaneous natural language. Due to the ad-hoc nature of conversational speech, such questions often include disfluencies or grammatical errors. We are unaware of any current work that is studying how those factors will affect the performance of ASR. Furthermore, previous studies were performed using off-the-shelf ASR systems as black boxes. In this study, we explored a learning-based language model adaptation approach to adapt the SRI system for recognizing spoken clinical questions.

METHOD

We evaluated the performance of the Medical and Generic versions of Nuance Dragon and the generic SRI Decipher system on recognition of spoken clinical questions. We then employed a language model approach for domain adaptation to improve ASR performance on spoken clinical questions. Since we could not change the models of the Nuance Dragon systems, we tested the effectiveness of the adapted model (SRI Adapted) based on the SRI Decipher system (SRI Gen). We tested the use of profile training to improve the performance of the Medical version of Nuance Dragon (Nuance Med). We analyzed the errors made to determine the effect of domain semantics.

ASR systems used

Nuance Dragon is a heavily used off-the-shelf ASR system, especially for dictation applications such as creating radiology reports. It has versions tuned for various genres. In this study, we employed the Medical and Generic versions (Nuance Med and Nuance Gen).

The SRI Decipher system used for all our experiments is a conversational speech-recognition system jointly developed by SRI and ICSI for the NIST Rich Transcription speech-recognition evaluation.[31] This system and its variants have shown state-of-the-art performance in the 2004, 2005, and 2006 NIST evaluations.

Language model adaptation for spoken clinical questions

Model adaptation is required to tune a basic set of models to specific acoustic and lexical characteristics (eg, to accommodate different types of acoustic conditions or semantic requests) or to subsets of speakers (eg, a different age group). Typically there are several different submodels in an ASR system, such as acoustic model, language model and phonetic model. To limit the scope of this study, we only investigated adaptation of the language model (LM) in this paper. The recognizer in the SRI system uses Kneser–Ney-smoothed[32] bigram, trigram, and 4-gram LMs at various stages of decoding. The baseline LMs are constructed by static interpolation of models from different sources, including meeting transcripts, topical telephone conversations, web data, and news; details can be found in Ozgur and Andreas.[33] When adapting the LMs using the strategies described below, all versions of the LMs used in the recognition system (bigram, trigram, 4-gram) were adapted similarly.

Two popular approaches for LM adaptation are model interpolation and count mixing.[34] In model interpolation, an out-of-domain model Θ_{OOD} is interpolated with an in-domain model Θ_{ID} to form an adapted model Θ:

$$P_{\Theta}(w_i|h_i) = \alpha P_{\Theta_{OOD}}(w_i|h_i) + (1-\alpha) P_{\Theta_{ID}}(w_i|h_i)$$

where $P(w_i|h_i)$ is the probability of the current word w_i given the history of $n-1$ words, h_i, in an n-gram LM (Θ, Θ_{OOD}, or

Θ_{ID}). α is a weight in [0,1], controlling the influence of the out-of-domain data on the final model and is usually optimized on a development set. Another approach to LM adaptation is count mixing, where the n-gram counts from all sources are summed, often after applying source-specific weights.

In this study, we employed supervised LM adaptation, that is, we used the manual transcription of clinical questions as collected by NLM (http://clinques.nlm.nih.gov), excluding the set used for ASR experiments. The interpolation weight was not tuned, and the default value of 0.5 was used. The tuning of parameters usually results in better speech-recognition performance, but our goal in this study was to determine the potential of the approach, and tuning is left for future work. As this set introduced more than 2000 new medical terms that are missing in the generic LM, we used a heuristic-based pronunciation estimation system to augment the pronunciation dictionary. Note that, while the medical terms are long enough to facilitate recognition, pronunciations listed in the dictionary are critical, and the estimation is not an error-free process.

EXPERIMENT AND RESULTS

Our goal is to evaluate the existing well-known speech-recognition systems on spoken clinical questions. We also evaluated adaptation methods and the effects of domain semantics. In order to eliminate the effect of acoustic differences that were present in all the previous comparative studies, we used the same set-up to record all the spoken questions to better compare the performance of all systems (as is typically done in speech-recognition research). Furthermore, we report the sum of substitution, insertion, and deletion errors (referred to collectively as word error rates (WER), an established metric for evaluating speech-recognition systems), instead of just analyzing substitution and deletion, as in some previous studies—for example, Devine *et al*.[17]

Data collection and experimental setup

In this study, we randomly selected 180 questions from the 4654 clinical questions used in Yu and Cao.[35] In the clinical-question data collection, all the clinical questions were deidentified to preserve the confidentiality of the physician asking the question and the patient or patients referred to in the question. The clinical questions were not edited for typographical errors.

To create the spoken data needed for our experiment, with the approval of Institutional Review Board at University of Wisconsin-Milwaukee we recruited nine medical students in their second year of medical school at the Medical College of Wisconsin. All of them were native English speakers and used English as their primary language. The subjects were each assigned 20 questions for the speech recording, and for each question, two types of speech were recorded: speech that was read verbatim (referred to as 'Read') and speech in which the question was asked in the subjects' own words spontaneously (referred to as 'Spoken'). In return, we offered each participant a small gift (~$15) for their participation.

Out of the 180 clinical questions that were recorded, we systematically evaluated the performance of five speech-recognition systems on 120 questions, based on the WERs established by the NIST Sclite standard scoring script. The five systems consisted of the four aforementioned systems (ie, Nuance Generic, Nuance Medical, SRI Generic, SRI Adapted) and a combined system that merged the results from the four systems based on a majority voting method, which aligns all four ASR outputs and selects the words that are included in the output of a majority of recognizers. We then used the recording data of the remaining 60 questions to investigate the effects of profile training on the Nuance Medical system. In our evaluation, the original clinical questions were used as references for the 'Read' setting, and one native English speaker manually transcribed the subjects' own versions of the questions as references for the 'Spoken' setting, where vocal pauses and stammering that were present in the spontaneous speech were included. For the two data sets, the average number of words per clinical question was 25.2 for the set of 120 questions and 38.1 for the set of 60 questions; the average number of sentences per clinical question was 1.4 and 2.0, respectively.

Speech-recognition results

Table 2 presents the WERs of the five systems on both the 'Read' and 'Spoken' setting of the 120 clinical questions. The Nuance Med system was shown to outperform the Nuance Gen system by improving the overall WER from 68.1% to 67.4%, but the performance dropped slightly (from 67.3% to 67.6% not statistically significant) for the 'Read' setting. The SRI systems achieved WER of 41.5% (SRI Gen), and our language model adaptation approach significantly improved the recognition performance leading to WER of 26.7% (SRI Adapted). But the performance of the combined system did not show any improvement as had been expected; on the contrary, the performance was slightly worse (WER of 28.2%) compared to the SRI Adapted system (WER of 26.7%).

We further analyzed the performance among the five systems described above in terms of deletion error rate, insertion error rate, and substitution error rate, respectively. Our results show that there is no statistical difference in WER between the Nuance Med and the Nuance Gen on spoken clinical questions, as shown in table 3. We found that although the overall performance of both SRI Gen and SRI Adapted was good, as shown in table 2, they did not perform well in terms of insertion rate, with 10.8%/9.8% and 6.1%/6.8% compared to other systems at 1.7%/2.3% and 1.3%/1.9%, respectively ($p<0.001$). The combined system outperforms the Nuance Med systems with respect to deletion rate (8.3%/8.8% vs 27.4%/26.5%) and the SRI Adapted systems for insertion rate (4.1%/4.6% vs 6.1%/6.8%), achieving the best substitution rate of 13.8% and 17.1 on the Read and Spoken setting, respectively. However, the combined system did not improve the overall performance, as shown in table 2. As for the comparison between the 'Read' and 'Spoken' settings, different systems presented mixed results based on differences in error rate.

Table 2 Speech-recognition performance for different automatic speech-recognition systems

System	Read (%)	Spoken (%)	Total (%)
Nuance Gen	67.3	69.1	68.1
Nuance Med	67.6	67.2	67.4
SRI Gen	40.7***	42.4***	41.5***
SRI Adapted	24.5***	29.3***	26.7***
Combined	26.2	30.5*	28.2
Nuance Med	37.8	51.1	41.9
Nuance Med w/Profile	27.0***	37.3***	30.1***

The upper part presents the performance comparison on a subset of 120 questions, including Nuance Dragon v.10.1 generic (Nuance Gen) and medical (Nuance Med) dictation systems, the SRI Decipher generic (SRI Gen) and adapted (SRI Adapted) conversational speech-recognition systems, and a system combining results from the above four systems (Combined). The bottom part presents the effects of Nuance profile training (Nuance Med vs Nuance Med w/Profile) on a different subset of 60 questions. Significant results (compared to the immediately above row) based on the t test are indicated by *$p<0.1$; **$p<0.01$; ***$p<0.001$.

Table 3 Further performance analysis among five systems

	Deletion rate		Insertion rate		Substitution rate	
	Read (%)	Spoken (%)	Read (%)	Spoken (%)	Read (%)	Spoken (%)
Nuance Gen	26.9	26.4	1.7	2.3	38.7	40.4
Nuance Med	27.4	26.5	1.3	1.9	38.9	38.8
SRI Gen	5.7***	7.3***	10.8***	9.8***	24.2***	25.3***
SRI Adapted	3.5*	3.4***	6.1***	6.8**	14.9***	19.1***
Combined	8.3***	8.8***	4.1***	4.6***	13.8	17.1**

Significant results (compared to the immediately above row) based on the t test are indicated by *p<0.1; **p<0.01; ***p<0.001.

Effects of profile training on nuance systems

We also investigated the effects of the profile training feature available on the Nuance Med system by randomly selecting three medical students from our recording subjects (corresponding to the 60 questions we recorded) and creating profiles for each of them using the standard procedure of the Nuance Dragon system. Since the system only allows for one vocabulary to be used for each profile, we chose Family Medicine for our study because the questions we evaluated were posed by family physicians. Thus, we were only able to compare the results from Nuance Medical (with vs without profile training), as shown in table 2. We can see that profile training yields a significant performance improvement for Nuance Medical, with the overall error rate being reduced from 41.9% to 30.1%, the error rate for the 'Read' setting from 37.8% to 27%, and the error rate for the 'Spoken' setting from 51.1% to 37.3%.

Error analysis by Unified Medical Language System ontology mapping

We took a step further to gain a better understanding of the effect of domain semantics on recognition errors. To that end, we developed a system to analyze recognition errors semantically by mapping all the words in clinical questions to the Unified Medical Language System (UMLS) Metathesaurus[36] using MMTX (http://www.nlm.nih.gov/research/umls/implementation_resources/mmtx.html) and defining three metrics to measure the extent to which the generated errors in the recognition process relate to medical concepts, the corresponding semantic types, and medical terms. The first was conceptErrorR, which is the percentage of UMLS concepts in the original questions that are related to recognition errors; the second was semTypeErrorR, which is the percentage of UMLS semantic types in the original questions that are related to recognition errors; and the third was medTermErrorR, which is the percentage of medical terms in the original questions that are related to recognition errors. The results are shown in table 4.

As anticipated, we observe error patterns at the semantic level similar to those at word level (WER). We can see that 68–78% of the medical terms in the clinical questions were incorrectly recognized by the Nuance systems, and 35–54% for the SRI systems. For the semantic type error rate and concept error rate, Nuance systems performed at 62–70% and 56–68%, and SRI system performed at 24–40% and 19–36%, respectively. The combined system was competitive in its ability to recognize domain semantics, as shown in the last row, but its performance was still worse than the SRI Adapted system. In addition, we found that the semantic level performance on the 'Spoken' setting was consistently lower than that of the 'Read' setting (comparing the two columns for each metric in table 4).

We observed that a total of 140 ('Read') and 139 ('Spoken') semantic types are involved in the 120 clinical questions, but the top 50 frequent ones account for 94.9% ('Read') and 95.0% ('Spoken') of all the medical terms that can be mapped to the UMLS. We thus conducted another analysis of error patterns focusing on only those 50 semantic types, finding lower error rate in terms of recognizing medical terms. In addition, we explored the relationship between semantic type frequency and corresponding term error rate for the top 50 semantic types. The results show a positive correlation on the Nuance system (except the 'Read' setting for Nuance Med) and negative correlation on the SRI and Combined system.

DISCUSSION

We found that none of the existing ASR systems performed well on spoken clinical questions, possibly because they were tuned for other applications. The language-model-based domain adaptation to the SRI Decipher system was quite successful, however, and the SRI Adapted system yielded the best total error rate of 26.7% (significant drop from 41.5% of the SRI Gen system), as shown in table 2. This indicates the importance of contextual information for speech recognition in the clinical domain. A good example is provided by the medical term 'Paget's disease.' Even though the second word in this term is covered by the generic model, the first word was not in its vocabulary and was therefore misrecognized by the SRI Gen model. However, the adapted LM can recognize these two words as a collocation, which allows for easier recognition of the first word based on the recognition of the second word. Domain-specific adaptation is crucial if existing generic systems are to be applied in the clinical domain.

In addition to language model adaptation, speaker-specific training appears to be as helpful for speech recognition on spoken clinical questions as it is in other domains as illustrated by the 28.2% drop (from 41.9% to 30.1%) in the total error rate

Table 4 Error analysis on the semantic level (macro average over questions)

	medTermErrorR		semTypeErrorR		conceptErrorR	
	Read (%)	Spoken (%)	Read (%)	Spoken (%)	Read (%)	Spoken (%)
Nuance Gen	76.0	77.9	68.6	70.5	67.5	68.4
Nuance Med	67.5**	69.9***	59.7***	61.6***	55.9*	61.2*
SRI Gen	51.2***	53.9***	39.9***	41.6***	33.9***	36.4***
SRI Adapted	35.4***	38.9***	24.3***	27.5***	18.9***	23.9***
Combined	37.9	43.6**	28.8**	34.4***	21.7*	28.8**

Significant results (compared to the immediately above row) based on the t test are indicated by *p<0.1; **p<0.01; ***p<0.001.

of Nuance Med after speaker-specific training, as shown in table 2. The results suggest that speaker specific training, which can be thought of as one acoustic model adaptation, can also help improve the ASR performance of spoken clinical questions. Therefore, we speculate that the SRI Adapted system has the potential for further improvement if we integrate a profile training feature and other adaptation techniques.

As shown in table 3, the Nuance systems achieved better performance on insertion rate than deletion rate. We speculate that dictation is more prone to deletion errors because dictation speech tends to be faster and more fluent than conversational speech, which involves more disfluencies, such as fillers, pauses, and stammering, which could promote extra insertions. The language-model adaptation we explored on the SRI system can capture such linguistic characteristics, yielding a better performance. Our results show that different systems vary in ways that affect different error rate metrics. In addition, we observed that combining the results from different systems tends to yield a compromised performance, but overall it did not improve the recognition performance due to the overwhelming effect of some systems.

Based on the semantic-level error analysis shown in table 4, we observed that the medical term error rate (medTermErrorR) was consistently higher than the semantic type error rate 'semTypeErrorR' and concept error rate 'concepErrorR' for all the systems. This indicates that some incorrectly recognized medical terms can still be mapped onto the correct semantic types and concepts, which we believe might alleviate the adverse effects of word-level recognition errors in real applications. For this semantic-level analysis, we focused on the errors relating to the original domain semantics (corresponding to deletion and substitution rate at the word level) regardless of what improper semantic information was inserted. When comparing the results of tables 3, 4, we noticed that the medical term error rate (medTermErrorR) was higher than the sum of the deletion and substitution rate (eg, 35.4%/38.9% vs 18.4%/22.5% for the SRI Adapted system), demonstrating additional challenges faced by an ASR system interpreting clinical question speech in comparison to other general domains. Note that the automatic mapping to the UMLS during our semantic analysis was not perfect, and additional research is needed to validate our findings.

This study focused on the evaluation of ASR systems on clinical question speech in a laboratory setting. The better performance on WER and deletion/substitution/insertion rates achieved by the SRI Adapted system will not necessarily transfer to real-life applications. For example, errors in recognizing function words would not be as important as those in recognizing content words in the QA task, as the presence or absence of function words may not change the QA performance. Nevertheless, our findings provide a foundation for further improving ASR performance in a clinical spoken QA system.

CONCLUSION

The results of this study show that ASR systems do not perform well on spoken clinical questions when applied without domain adaptation or speaker-specific training. Learning-based language model adaptation and speaker-specific training each can improve performance significantly. Using both in combination may further improve performance.

The language-model-based adaptation explored in this study was simple but very effective, which suggests that ASR performance on spoken clinical questions may be further improved by employing more sophisticated language-based adaptation models. We intend to investigate more sophisticated language models, as well as adaptation methods for other components, such as the acoustic and phonetic models in the ASR system. For example, we plan to integrate profile training in the SRI Adapted system and develop a more systematically combined system for our future work. More research and experiments on larger data sets are needed to validate our findings. We also intend to build a publicly available clinical QA systems and to evaluate different ASR systems in real-world settings.

Acknowledgments The SRI conversational speech-recognition system is not off the shelf and requires support for obtaining any results. This is actually the case for many speech-recognition systems built in the industrial research labs and other academic institutions. This is why, in the medical transcription literature, there is no fair comparison between these systems and off the shelf systems. GT, from SRI, is an established researcher working for the SRI Research Labs (and formerly at AT&T Research Labs) well known in the speech-processing community for his work on conversational speech recognition and understanding, and has many academic awards. As the SRI speech-recognition system is not available for any commercial use or licencing at this point, there is no commercial conflict of interest, and this effort has only aimed at providing a fair technical performance comparison long needed by the medical informatics community, which has less insider knowledge about such speech-recognition systems. The coauthors GT and DH-T have provided very useful feedback throughout this study from the selection of the audio recorder to experimental setup to data preprocessing to the use of the SRI speech-recognition systems. The authors thank AM Kruse, for collecting the speech data of clinical questions used in this study, writing up its protocol, and performing statistical T tests. The authors thank L Antieau, for proofreading the initial manuscript, C Kahn and A Bennett, for their helpful advice and discussion, and all the medical students who participated in the study.

Funding The authors acknowledge support from the National Institute of Health (NIH) grant number 5R01LM009836 and a seed funding from University of Wisconsin-Milwaukee.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Ely JW,** Osheroff JA, Ebell MH, *et al*. Analysis of questions asked by family doctors regarding patient care. *BMJ* 1999;**319**:358—61.
2. **Timpka T,** Arborelius E. The GP's dilemmas: a study of knowledge need and use during health care consultations. *Methods Inf Med* 1990;**29**:23—9.
3. **Bergus GR,** Randall CS, Sinift SD, *et al*. Does the structure of clinical questions affect the outcome of curbside consultations with specialty colleagues? *Arch Fam Med* 2000;**9**:541—7.
4. **Ely JW,** Burch RJ, Vinson DC. The information needs of family physicians: case-specific clinical questions. *J Fam Pract* 1992;**35**:265—9.
5. **Osheroff JA,** Forsythe DE, Buchanan BG, *et al*. Physicians' information needs: analysis of questions posed during clinical teaching. *Ann Intern Med* 1991;**114**:576—81.
6. **Covell DG,** Uman GC, Manning PR. Information needs in office practice: are they being met? *Ann Intern Med* 1985;**103**:596—9.
7. **Smith R.** What clinical information do doctors need? *BMJ* 1996;**313**:1062—8.
8. **David L.** Sackett: Evidence-based medicine. *Semin Perinatol* 1997;**21**:3—5.
9. **Gosling AS,** Westbrook JI. Allied health professionals' use of online evidence: a survey of 790 staff working in the Australian public hospital system. *Int J Med Inform* 2004;**73**:391—401.
10. **Westbrook JI,** Gosling AS, Coiera E. Do clinicians use online evidence to support patient care? A study of 55,000 clinicians. *J Am Med Inform Assoc* 2004;**11**:113—20.
11. **Westbrook JI,** Coiera EW, Gosling AS. Do online information retrieval systems help experienced clinicians answer clinical questions? *J Am Med Inform Assoc* 2005;**12**:315—21.
12. **Ely JW,** Osheroff JA, Ferguson KJ, *et al*. Lifelong self-directed learning using a computer database of clinical questions. *J Fam Pract* 1997;**45**:382—8.
13. **D'Alessandro DM,** Kreiter CD, Peterson MW. An evaluation of information-seeking behaviors of general pediatricians. *Pediatrics* 2004;**113**:64—9.
14. **Ely JW,** Osheroff JA, Chambliss ML, *et al*. Answering physicians' clinical questions: obstacles and potential solutions. *J Am Med Inform Assoc* 2005;**12**:217—24.
15. **Zafar A,** Overhage JM, McDonald CJ. Continuous speech recognition for clinicians. *J Am Med Inform Assoc* 1999;**6**:195—204.
16. **Zafar A,** Mamlin B, Perkins S, *et al*. A simple error classification system for understanding sources of error in automatic speech recognition and human transcription. *Int J Med Inform* 2004;**73**:719—30.
17. **Devine EG,** Gaehde SA, Curtis AC. Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports. *J Am Med Inform Assoc* 2000;**7**:462—8.

18. **Borowitz SM.** Computer-based speech recognition as an alternative to medical transcription. *J Am Med Inform Assoc* 2001;**8**:101—2.
19. **Mohr DN,** Turner DW, Pond GR, *et al*. Speech recognition as a transcription aid: a randomized comparison with standard transcription. *J Am Med Inform Assoc* 2003;**10**:85—93.
20. **Havstam C,** Buchholz M, Hartelius L. Speech recognition and dysarthria: a single subject study of two individuals with profound impairment of speech and motor control. *Logoped Phoniatr Vocol* 2003;**28**:81—90.
21. **Pezzullo JA,** Tung GA, Rogg JM, *et al*. Voice recognition dictation: radiologist as transcriptionist. *J Digit Imaging* 2008;**21**:384—9.
22. **Butzberger J,** Murveit H, Shriberg E, *et al*. Spontaneous speech effects in large vocabulary speech recognition applications. *Proceedings of the Workshop on Speech and Natural Language*. Harriman, NY: Association for Computational Linguistics, 1992:339—43.
23. **Rangarajan V,** Narayanan S. Analysis of disfluent repetitions in spontaneous speech recognition. *Proceedings of EUSIPCO*. 2006.
24. **Niesler T,** Willett D. Unsupervised language model adaptation for lecture speech transcription. *Interspeech*. 2002.
25. **Tur G,** Stolcke A. Unsupervised language model adaptation for meeting recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2007;**4**:IV-173—IV-176.
26. **Liu Y,** Liu F. Unsupervised language model adaptation via topic modeling based on named entity hypotheses. *Proceedings of ICASSP* 2008;**8**:4921—4.
27. **Mandel MA.** A commercial large-vocabulary discrete speech recognition system: DragonDictate. *Lang Speech* 1992;**35**:237—46.
28. **Posur W.** Experiences with a current speech recognition system in creating cardiology reports. *Herz* 2000;**25**:627—32.
29. **Issenman RM,** Jaffer IH. Use of voice recognition software in an outpatient pediatric specialty practice. *Pediatrics* 2004;**114**:e290—3.
30. **Chase HS,** Kaufman DR, Johnson SB, *et al*. Voice capture of medical residents' clinical information needs during an inpatient rotation. *J Am Med Inform Assoc* 2009;**16**:387—94.
31. **Stolcke A,** Anguera X, Boakye K, *et al*. *Further Progress in Meeting Recognition: the ICSI-SRI Spring 2005 Speech-to-Text Evaluation System. Vol. 3869*. LNCS, MLMI Workshop, 2005;**78**:463—475.
32. **Kneser R,** Ney H. Improved backing-off for M-gram language modeling. *ICASSP* 1995;**1**:181—4.
33. **Ozgur C,** Andreas S. *Language Modeling in the ICSISRI Spring 2005 Meeting Speech Recognition Evaluation System*. Berkeley, CA: International Computer Science Insititute, 2005.
34. **Bellegarda JR.** Statistical language model adaptation: review and perspectives. *Speech communication* 2004;**42**:93—108.
35. **Yu H,** Cao YG. Automatically extracting information needs from ad hoc clinical questions. *AMIA Annu Symp Proc* 2008:96—100.
36. **Humphreys BL,** Lindberg DA, Schoolman HM, *et al*. The Unified Medical Language System: an informatics research collaboration. *J Am Med Inform Assoc* 1998;**5**:1—11.