Research Paper ■

A Continuous-speech Interface to a Decision Support System: II. An Evaluation Using a Wizard-of-Oz Experimental Paradigm

WILLIAM M. DETMER, MD, SMADAR SHIFFMAN, MS, JEREMY C. WYATT, DM, MRCP, CHARLES P. FRIEDMAN, PHD, CHRISTOPHER D. LANE, LAWRENCE M. FAGAN, MD, PHD

Abstract **Objective**: Evaluate the performance of a continuous-speech interface to a decision support system.

Design: The authors performed a prospective evaluation of a speech interface that matches unconstrained utterances of physicians with controlled-vocabulary terms from Quick Medical Reference (QMR). The performance of the speech interface was assessed in two stages: in the real-time experiment, physician subjects viewed audiovisual stimuli intended to evoke clinical findings, spoke a description of each finding into the speech interface, and then chose from a list generated by the interface the QMR term that most closely matched the finding. Subjects believed that the speech recognizer decoded their utterances; in reality, a hidden experimenter typed utterances into the interface (Wizard-of-Oz experimental design). Later, the authors replayed the same utterances through the speech recognizer and measured how accurately utterances matched with appropriate QMR terms using the results of the real-time experiment as the "gold standard."

Measurements: The authors measured how accurately the speech-recognition system converted input utterances to text strings (recognition accuracy) and how accurately the speech interface matched input utterances to appropriate QMR terms (semantic accuracy).

Results: Overall recognition accuracy was less than 50%. However, using language-processing techniques that match keywords in recognized utterances to keywords in QMR terms, the semantic accuracy of the system was 81%.

Conclusions: Reasonable semantic accuracy was attained when language-processing techniques were used to accommodate for speech misrecognition. In addition, the Wizard-of-Oz experimental design offered many advantages for this evaluation. The authors believe that this technique may be useful to future evaluators of speech-input systems.

■ **J Am Med Informatics Assoc.** 1995;2:46–57.

The evaluation of medical computer systems is a difficult but important task.[1–4] The evaluation of user interfaces is especially difficult, because measurements of interface characteristics are often subjective and thus prone to observational bias. In addition, there are often many interface elements to study and those elements are frequently interdependent. Finally, standards by which to compare new interface

Affiliations of the authors: Section on Medical Informatics, Stanford University School of Medicine, Stanford, CA (WMD, SS, CDL, LMF); Biomedical Informatics Unit, Imperial Cancer Research Fund Laboratories, London, England (JCW); and Laboratory for Computing and Cognition, University of North Carolina School of Medicine, Chapel Hill, NC (CPF).

Supported by the National Library of Medicine under grants LM-04864 and LM-07033. Computer facilities were provided by the SUMEX-AIM project (LM-05208), by the CAMIS Resource (LM-05305), and through equipment loans from Speech Systems, Inc. Dr. Wyatt's work was funded by a UK Medical Research Council Traveling Fellowship.

Correspondence and reprints: William M. Detmer, MD, Section on Medical Informatics, Medical School Office Building X215, Stanford University School of Medicine, Stanford, CA 94305-5479.

Received for publication: 5/31/94; accepted for publication: 9/07/94.

components do not exist. Therefore, evaluators face at least three challenges when designing evaluation studies: 1) to control as many nonessential variables as possible and thus allow careful study of important variables; 2) to design experiments that separate the system's components so that they can be studied both individually and together; and 3) to design experiments that have an ideal standard against which to measure system performance.

In this article, we report the evaluation of a speech interface developed in our laboratory.[5,6] This interface facilitates entry of clinical findings into the decision support system Quick Medical Reference (QMR)*[7] by matching the spoken language of physicians with the controlled vocabulary of the decision support system. The interface uses a commercially available speech-recognition system to produce text strings from utterances and then a concept-matching approach to match text strings with QMR terms.

We conducted a Wizard-of-Oz experiment to assess the performance of the speech interface. Users interacted with the system as if they were controlling the interface with speech, while instead a hidden experimenter (the "wizard") typed every utterance into the interface. This experimental design had two purposes: 1) to measure the system's components both together and separately, and 2) to establish "gold standards"—approximations of ideal performance. These "gold standard" measurements could then be used as the basis for determining the accuracy of the system's speech-recognition and concept-matching components.

Background

Below we describe methods that have been used to evaluate speech interfaces, and we review how the Wizard-of-Oz paradigm has been used to design speech systems. We also describe the architecture of the speech-recognition interface to QMR.

Evaluation Techniques for Speech Interfaces

Speech-recognition interfaces have been evaluated along several dimensions. Some evaluations focus on comparisons between speech input and other input modalities,[8,9] while others focus on the quality of the speech recognition and the contribution of speech recognition to the accomplishment of an application task.[10–12]

Comparisons of speech interfaces with other input modalities focus primarily on task-completion time—i.e., how long it takes a user to complete a task using speech versus using a standard input modality such as a keyboard or a pointing device.[8,9] Task-completion time includes such factors as the response time of the speech-recognition equipment and the time users spend correcting errors of the speech recognizer. Although useful for comparison with other input modalities, this method of evaluation does not directly measure the quality of the speech recognition or the contribution of speech recognition to the application task.

The more traditional method of evaluating speech-recognition systems is to measure 1) how accurately the speech recognizer converts input utterances to text strings (*recognition accuracy*), and 2) how accurately the speech application translates input utterances to appropriate actions (*semantic accuracy*). Recognition accuracy measures the raw capability of the speech-recognition equipment, while semantic accuracy measures how well speech input supports application tasks. The relative importance of each of these measures depends on the application studied. If the output of the speech recognizer is not displayed to the user but instead is used by the application to perform a task, then the semantic accuracy is a more important measure of the system's performance.

On the other hand, if the user takes action based on the exact output of the speech recognizer, then the recognition accuracy may be the more important measure. The difference between the semantic accuracy and the recognition accuracy shows the contribution of higher-level application components, such as language-processing routines, that speech systems employ to overcome poor recognition accuracy.

Wizard-of-Oz Experiments for Design of Speech Interfaces

Wizard-of-Oz experiments have been used primarily to help in the design of speech-recognition systems.[13–15] The experimental setting requires that two computer terminals in physically distinct locations be linked to the same central processing unit and thus show the same view. Sitting at one terminal, the subject uses natural language to give commands to the system. Unseen by the subject, an experimenter sitting at the other terminal interprets the subject's requests and translates them into commands that the system can understand. The subject is therefore given the impression that the system can understand natural language.

*Quick Medical Reference and QMR are registered trademarks of the University of Pittsburgh.

The main purpose for these Wizard-of-Oz experiments has been to understand how users would speak to a computer that recognizes spoken input. For instance, Gould et al.[14] used the Wizard-of-Oz paradigm to study how users would interact with a "listening typewriter," while Issacs et al.[15] performed a Wizard-of-Oz experiment to study how users would enter data into a medical decision support system. The results of these experiments helped the developers understand the general functional requirements for speech systems in these domains, as well as how to handle specific problems, such as how to correct for misrecognized input.

In contrast to these previous studies, we used the Wizard-of-Oz paradigm not for the design of a speech-recognition interface but for the evaluation of a speech-recognition system's performance. Coutaz et al.[16] have suggested using this paradigm for similar evaluation experiments.

The Continuous-speech Interface to QMR

The program that we evaluated consists of two main components: a speech-recognition system, which converts audio signals into text strings, and a concept matcher, which matches the recognized text string with QMR terms. The speech-recognition system is composed of two subcomponents: off-the-shelf hardware and software from Speech Systems, Inc. and developer-created grammar that defined legal sentences for the domain. Further details of the system design are presented in the companion article in this issue.[5]

The physician follows three steps when using the continuous-speech interface. First, the physician selects a body part corresponding to the location of the clinical finding (e.g., abdomen). Second, the physician speaks the finding into the microphone. The speech recognizer first decodes the audio signal into a stream of subphonemes and then generates a text string using both a built-in dictionary and a developer-supplied grammar. The concept matcher then extracts keywords from the recognized text string and compares those keywords with keywords extracted from QMR terms. The program displays the result of the matching as a rank-ordered list of QMR terms. Third, the physician selects the term that most closely matches the intended finding.

In this experiment, we studied how the system performed with two types of grammars: 1) grammars generated manually by a developer who had medical training, and 2) grammars generated programmatically from a set of general language rules. We studied different methods for creating grammars to understand which grammar would support better recognition and semantic accuracy.

Design Considerations

An ideal strategy for evaluating a speech interface would be to place a fully functioning system in a real-world environment and study the use, performance, and perceived value of the system during normal work flow, as well as the impact of the system on patient care and outcomes.[2] However, for research prototypes this is impractical because such prototypes are not sufficiently developed to be readily adopted in normal work flow. Therefore, we set out to evaluate the speech interface in a simulated but realistic setting. In the design of the evaluation, we faced several major problems: 1) how to elicit natural speech from physicians without biasing what they said; 2) how to measure overall system accuracy, as well as the accuracy of the individual system components; 3) how to establish standards by which the accuracy of the system components could be measured; and 4) how best to compare the two grammars.

Eliciting Expressions from Subjects without Introducing Bias

To evaluate how well the continuous-speech interface might perform in a real-world setting, we first needed to elicit from physicians unbiased, natural-language expressions. One method for eliciting such expressions would be to have physicians examine patients with abnormal clinical findings and then ask the physicians to speak those findings into the interface. This method would mirror closely how the system would be used in clinical practice. The disadvantage of this method is the time, cost, and complexity associated with standardizing the experimental setup and recruiting patients and physician subjects.

Another method for eliciting expressions would be to give physicians findings from dictated progress notes and ask them to speak a description of the findings into the interface using their own words. This method would elicit expressions in an efficient manner and simplify the experimental setup. However, presenting text descriptions to physicians could potentially bias them toward speaking words and phrases from the text descriptions.

A final method for eliciting expressions would be to use audiovisual stimuli. For instance, we could evoke the finding *right lower quadrant tenderness* using a diagram of the abdomen with the right lower quadrant highlighted and the sound "ouch!," which would

play when the user touched the highlighted area. Presenting audiovisual stimuli would allow physicians to form a clinical concept in their minds without verbal bias and then to describe the concept to the interface in their natural language. The disadvantage of this approach is the difficulty of assembling multimedia material that would reproducibly elicit a clinical finding.

Studying Overall and Component Accuracy

We wished to measure the semantic accuracy of the speech interface and understand the relative contributions of the speech-recognition and concept-matching components. To do this, our experimental design should allow measurement of the system's components both separately and together. Especially difficult would be measuring the accuracy of the concept matcher alone. Normally, concept-matching accuracy would be measured by how well input utterances matched with QMR terms. However, if poor speech recognition distorted the input utterance before it reached the concept matcher, then the measurement of concept-matching accuracy would also be distorted. Therefore, our experimental design required that in the real-time experiment the concept matcher receive undistorted input, such as a human transcription of each utterance.

Establishing "Gold Standards"

To measure the accuracy of the speech-interface program and its components, we needed to establish "gold standards." Particularly difficult was establishing the "gold standard" for concept matching. The main difficulty was deciding who should judge how well an utterance matched with the chosen QMR term. For instance, how well does the utterance "right lower quadrant pain" match with the QMR term *abdomen tenderness right lower quadrant*? One option would be to assemble a panel of physicians to review pairs of input utterances and vocabulary terms and to judge their semantic proximity. The advantage of this approach would be that the standards for judging proximity would be applied uniformly across all experimental subjects and all utterance–term pairs. The

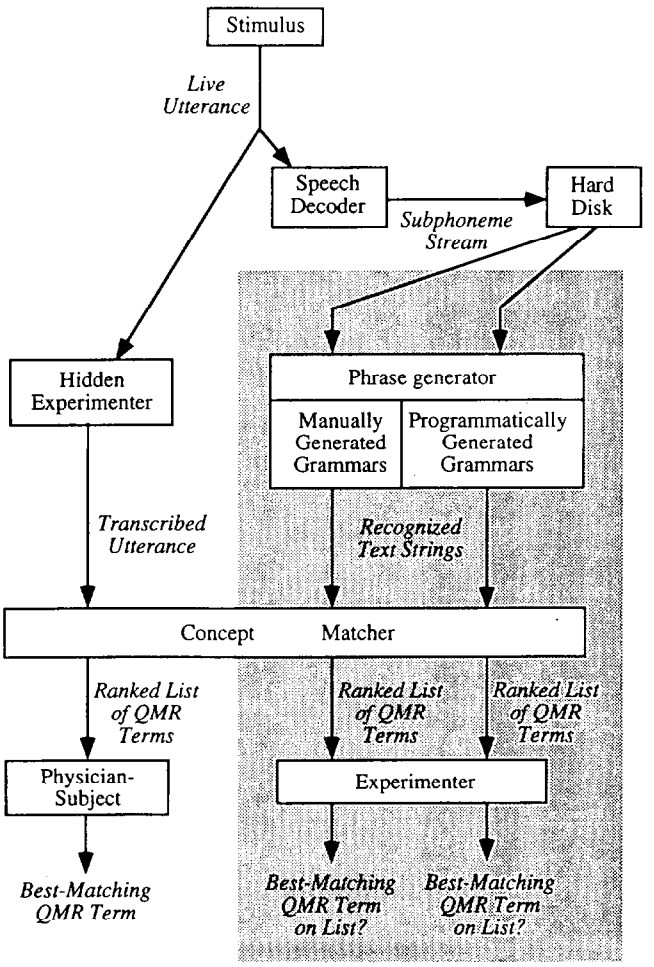

Figure 1 Overview of the evaluation experiment. During the real-time experiment (area outside the shaded rectangle), the subject's utterance was processed in parallel by a hidden experimenter and the speech-recognition system: the experimenter transcribed the utterance into the interface, and the speech-recognition system decoded the utterance to a subphoneme stream and stored the stream on disk. The transcribed utterance was processed by the concept matcher and generated a ranked list of Quick Medical Reference (QMR) terms; the subject then chose the closest matching term and identified how close it matched to the utterance. At a later time (area inside the shaded rectangle), the stored subphoneme streams were further processed by the speech recognizer twice, once using manually generated grammars and then using programmatically generated grammars. Each resulting recognized text string was processed by the concept matcher and generated a ranked list of QMR terms. The semantic accuracy of the speech interface was measured by observing whether the QMR term chosen in the real-time experiment (the "gold standard") appeared on the list when the utterance was run through both the speech recognizer and the concept matcher.

disadvantage of this method is that experts would lack contextual information that might help them estimate the intentions of subjects.

Another way to judge the semantic proximity of utterance–term pairs would be to ask experimental subjects to judge the degree to which their own input utterances matched with a QMR term. This method would be efficient and would allow users of the system to judge for themselves the accuracy of the matching. The disadvantage of this approach is interrater variability: subjects given the same utterance–term pair might have considerably different thoughts on how well the utterance matched with the controlled-vocabulary term.

Comparing the Grammars

To compare how the manually generated and programmatically generated grammars differed in their support of speech recognition, we needed to measure speech recognition using similar or identical inputs. One approach would be to ask physicians to speak the same phrase into the computer twice; the first utterance could be processed with one grammar and the second utterance with the other grammar. The disadvantage of this approach is that although the two utterances would contain identical words and word orders, they would likely contain different volumes, pauses, and inflections.

A more desirable strategy would be to compare the grammars using identical inputs. However, because our speech-recognition system could not process in parallel one utterance with two grammars, we needed a way to store utterances without degrading their acoustic quality and later process them serially using each grammar. Since the speech-recognition system decodes audio signals to a stream of subphonemes before generating a text string using a grammar, we could store the subphoneme stream during the real-time experiment and later generate phrases using each grammar.

Experimental Design

We used audiovisual stimuli and a Wizard-of-Oz experimental design to meet the requirements outlined above. An overview of the experiment is shown in Figure 1. We presented 20 audiovisual stimuli that represented QMR findings to 20 physician subjects and asked the physicians to speak those findings into the interface using their own words. During the real-time running of the experiment, subjects entered clinical findings using spoken phrases and then chose the QMR term that best matched the intended finding. Users believed that the list of QMR terms was generated by the speech-interface program when, in fact, the list was generated by a hidden experimenter typing the utterance into the concept matcher. This design allowed the concept matcher to receive as input not the output from the speech recognizer but instead a human transcription of the utterance. Thus, we could measure the accuracy of concept matching without contamination from speech misrecognition.

Later, we analyzed the accuracy of the entire system by processing the utterances made during the experiment twice, once using manually generated

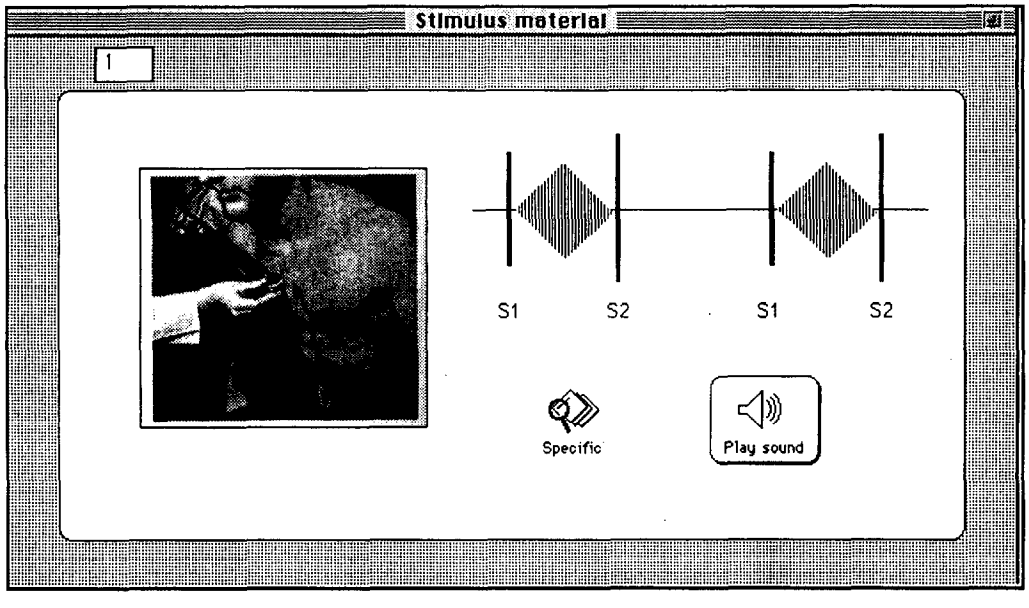

Figure 2 Example of a stimulus used to evoke the clinical concept represented by the Quick Medical Reference (QMR) term *heart murmur systolic ejection second right interspace*. The stimulus shows a stethoscope placed on a patient's chest and a representation of the cardiac cycle with a diamond-shaped murmur. Clicking the "Play sound" button produces an audio clip of a systolic murmur. The "Specific" icon signals the subject to "be as specific as you can" when describing the finding.

grammars and then using programmatically generated grammars. This method allowed us to measure how well the interface would have performed had we used the speech recognizer and a particular grammar in the real-time experiment.

Subjects

We used an invitation letter distributed by electronic mail and word of mouth to recruit 20 physicians from our medical center. Subjects were eligible to participate if they had completed medical school and internship, had no detailed knowledge of QMR or speech-recognition technology, and used English as a first language.

Warm-up

Although the continuous-speech recognition system was speaker-independent (i.e., it was capable of recognizing speech from different users without training), it did require a warm-up period to make adjustments for the volume of each speaker's voice. A warm-up period was also necessary to teach subjects how to use the speech apparatus: how to activate the speech-recognition system by pressing a button and coordinate this with speaking into a head-mounted microphone. Lastly, the warm-up period was necessary to give subjects confidence that the computer could recognize their utterances.

Because we did not want to bias what the subjects would say during the experiment, we had the subjects speak nonmedical phrases during the warm-up period. One of us (CDL) created a small speech application that recognized lines from a familiar children's book. The application recognized only a small number of phrases and thus had recognition accuracy approaching 100%. Subjects spoke supplied phrases into the application and observed near-instantaneous output from the speech recognizer. Subjects could use this feedback to confirm that a phrase was recognized or to reenter an utterance if it was misrecognized. Once subjects had gained confidence with the speech apparatus and the speech recognizer was consistently recognizing their utterances, we terminated the warm-up. The average warm-up time was five minutes.

Presentation of Audiovisual Stimuli

To minimize bias that might be introduced by textual material, we presented to subjects clinical findings that were represented predominantly by diagrams, images, and sounds. Stimuli were designed to evoke from physicians abnormal physical examination findings that are represented in the QMR knowledge base. To eliminate bias due to learning and fatigue, the stimulus material was presented to each subject in random order.

Table 1 ■

Characteristics of the Subjects of This Study

Characteristic	Value
Mean length of time since graduation from medical school	7.95 years
Male	(17/20) 85%
Completed residency	(18/20) 90%
Medical specialty	
Internal medicine	(15/20) 75%
Surgery	(2/20) 10%
Radiology	(1/20) 5%
Emergency medicine	(1/20) 5%
Internship only	(1/20) 5%

One of us (JCW), who had no knowledge of the speech interface or how the grammars were created, developed the stimulus material.[17] He was first given a list of all the QMR terms describing abnormal physical findings localized in the neck, back, breast, heart, chest, abdomen, and vasculature. Only QMR terms that contained at least one word found in the speech system's dictionary were included. The designer randomized these 116 terms and, starting with the first, selected 20 terms he considered most feasible to communicate using diagrams, images, and sounds. He rejected 17 QMR terms that he believed could not be communicated reliably, such as *abdomen flank heavy*, or *neck muscle flaccid*.

The designer assembled the audiovisual stimuli using both images from books about physical examination, medical atlases, and slide collections and clips from audio libraries (Fig. 2). When no suitable image or audio clip was available, the designer diagrammed the location of the abnormal physical finding using a graphics program. Most diagrams consisted of a body chart with the abnormality outlined or shaded.

Six pilot physician subjects, who were not subjects in the main experiment, reviewed the test stimuli and were asked to name the clinical finding. Stimuli that evoked too many different findings were improved by altering or adding diagrams, images, or audio. In 14 of the 20 findings, pilot subjects found it difficult to gauge what level of detail was portrayed. For instance, a diagram showing *splenomegaly moderate* might be described as an abdominal mass, a left upper quadrant mass, or splenomegaly. Icons were added to denote "be as specific as you can" or "this is a general finding."

Pilot studies showed that only four physical findings could be evoked reliably by visual stimuli alone, while another six findings required both visual and auditory stimuli. For the remaining findings, text was added to describe either a typical patient in whom the finding might be observed (e.g., for *splenomegaly moderate*: "Diagnosis: chronic malaria") or a procedure that would elicit that finding (e.g., "Procedure: palpation"). When text was added to the stimulus material, the designer used few words and was careful not to name the abnormality or the anatomic site.[17]

Measurement of Speech-recognition and Concept-matching Accuracy

Below we describe how we measured the accuracy of concept matching alone, of speech recognition alone, and of speech recognition and concept matching together.

Concept-matching accuracy. During the real-time experiment, the concept matcher received as input a human transcription of the utterance, not the output from the speech recognizer. If we assumed that the human transcription of the utterance was highly accurate, then we could measure the accuracy of the concept matcher alone by asking subjects to judge the degree of correspondence between their utterances and the QMR terms that they chose. Subjects could grade the match between utterance and controlled-vocabulary term as "exact," "general, but not exact," or "not a match."

Because the human experimenter could have made errors transcribing utterances, we listened to audio tapes recorded during the experiment and analyzed transcription errors. We observed how many transcriptions contained errors and whether the errors were misspellings or were word substitutions, deletions, or insertions. To understand whether transcription errors distorted the output of the concept matcher, we also observed whether errors altered the set of terms on the QMR list.

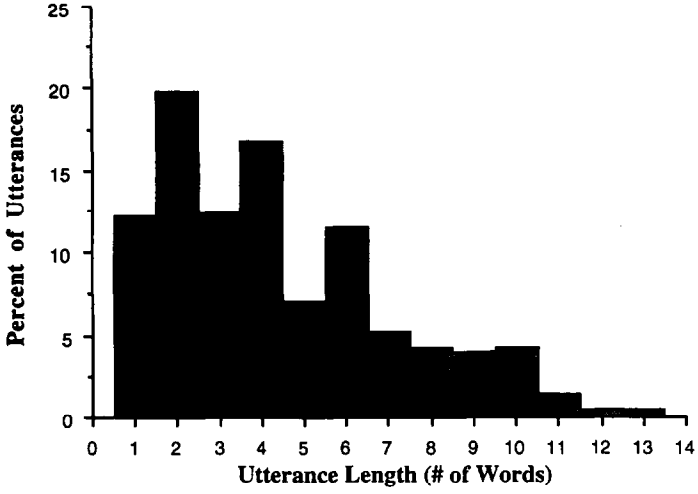

Figure 3 Frequency distribution of utterance length.

Speech-recognition accuracy. We measured speech-recognition accuracy by comparing the output of the speech-recognition system with the human transcription of the utterance. There were several ways we could have calculated speech-recognition accuracy. In general, we should give a positive score for words appearing in both the transcribed and the recognized phrases and a negative score for words that are inserted, deleted, or substituted in the recognized phrase. The most common approach for evaluating accuracy is to align the recognized string against the transcribed utterance using a dynamic programming algorithm and then count the words that are correct, substitutions, deletions, and insertions. The recognition accuracy can then be calculated using a formula.[18]

For our evaluation, we used a modification of this approach. First, we eliminated from the transcribed and recognized phrases all stop words (e.g., *the, an, of*) because we wished to focus on recognition accuracy of keywords. We then used the following formula to calculate the recognition accuracy from the transcribed and recognized phrases:

$$\text{Recognition accuracy (\%)} = \frac{\text{Number of words in common}}{\text{Number of unique words}} \times 100.$$

In this formula, the numerator shows the number of words appearing in both phrases (the intersection), while the denominator shows the number of unique words in the two phrases combined (the union). If the transcribed and recognized phrases are exactly the same (e.g., both are "right lower quadrant tenderness"), the numerator and denominator will be equal and the recognition accuracy will be 100%. If the recognized phrase contains insertions, deletions, or substitutions, the numerator will be small relative to the denominator and the recognition accuracy will be lower. For instance, if the transcribed utterance were "liver tenderness" and the recognized text string were "severe tenderness," there would be one word in common and three unique words. Therefore, the recognition accuracy would be 33%.

We chose this formula for computing recognition accuracy because of its simplicity. In addition, we had observed that an equivalent form of this formula, the Hooper equation,[19] has been used extensively in the literature to measure the degree of similarity between index terms assigned by different indexers.

Accuracy of speech recognition and concept matching together. We measured how well the interface program would have worked had we used speech recognition in real-time. We processed the subphoneme stream from each utterance with a particular grammar, and ran the resulting recognized text string through the concept matcher to create a ranked list of QMR terms. We measured semantic accuracy by observing whether the QMR term that the subject had chosen during the experiment (the "gold standard") appeared on the list of terms generated when the utterance was run through both the speech recognizer and the concept matcher.

Results

Subjects

Characteristics of the 20 subjects are shown in Table 1. All subjects were native English speakers; two were from English-speaking countries outside the United States.

Utterance Characteristics

With 20 subjects each speaking 20 utterances, we had a total of 400 utterances for analysis. Utterance length ranged from 1 to 13 words, with a mean of 4.39 words (Fig. 3). A typical one-word utterance was "bradycardia" and a representative long utterance was "a two out of six systolic murmur in the right upper sternal border." Analysis of variance showed that subjects differed significantly in the number of words they used to describe a clinical finding ($F = 9.624$, $p < 0.0001$). However, we found no significant correlation between utterance length and gender, specialty, or years since graduation from medical school.

Concept-matching Accuracy

Of the 400 utterance–term matches generated in the experiment, subjects judged 261 (65%) to be exact, 96 (24%) to be general, but not exact, and 43 (11%) to be not a match. Combining exact and general matches, 89% of the utterances in the experiment matched to a QMR term in at least a general way. Table 2 shows examples of matches that were classified in the three different categories. We found no correlation between the degree of matching and the subject's gender, specialty, or years since graduation from medical school, nor between the degree of match and the order in which the clinical finding was presented.

Most matching errors occurred with just a few of the QMR terms. For example, of utterance–term pairs that subjects judged as not a match, 34 of 43 (79%) were produced by 4 of 20 QMR findings. Of these findings, one (*breast gynecomastia*) accounted for 12 of 43 (28%) of the errors because the crucial keyword "gynecomastia" was accidentally omitted from the dictionary used by the interface program. Although the speech recognizer correctly recognized the word gynecomastia, the concept matcher did not match the word with the QMR term. The other findings caused errors for two reasons: 1) there was ambiguity in either the stimulus material that represented the findings or the QMR term for which the stimulus material was created, or 2) keywords extracted from the recognized string did not point to synonyms that would allow matching with the appropriate QMR term.

Table 2 ■

Examples of How Subjects Rated the Match between Their Utterances and the Quick Medical Reference (QMR) Terms That They Chose

Degree of Match	Utterance	QMR Term Chosen
Exact	"Right lower quadrant tenderness"	Abdomen tenderness right lower quadrant
General, but not exact	"Right-sided neck bruit"	Artery carotid systolic bruit
Not a match	"Massive ascites"	None*

*QMR terms from which subjects could choose were 1) splenomegaly massive, 2) splenomegaly moderate, and 3) liver enlarged moderate.

Ambiguity of either the stimulus material or the QMR term accounted for 17 of 43 errors (40%). For instance, the stimulus for *abdominal flank bulging bilaterally* generated utterances that included "protuberant abdomen," "bloated abdomen," "distended abdomen," "ascites," and "abdominal birthmark." Even though the intended QMR term appeared on the list in 15 of 20 cases, only six subjects chose this QMR term as the closest match. In addition, only one of these six subjects rated the degree of match as exact. Further analysis of the errors generated by the stimulus material is presented elsewhere.[17]

Lack of appropriate synonym pointers was the cause for the remaining 14 errors (38%). For instance, utterances such as "physical examination shows a large liver" and "enlarged liver" did not match with the QMR term *hepatomegaly* because the canonical form for *liver* did not include pointers to *hepatomegaly*.

To understand how transcription errors may have distorted the output of the concept matcher and affected its accuracy, we compared the experimenter's

Table 3 ■

Recognition Accuracy, Semantic Accuracy, and Quick Medical Record (QMR) List Length for the Two Grammars

Grammar	Recognition Accuracy ± SEM (%)	Semantic Accuracy (%)	Mean QMR List Length ± SEM
Manually generated	47 ± 1.7*	81†	8.44 ± 0.341
Programmatically generated	39 ± 1.8	74	10.89 ± 0.360‡

*t-test, $p < 0.0001$.
†χ^2, $p < 0.0001$.
‡t-test, $p < 0.0001$.

Table 4 ■

Examples of Speech-recognition Errors

Utterance	Grammar Used by Speech Recognizer	
	Manually Generated	Programmatically Generated
"He has a crescendo decrescendo diastolic murmur"	A harsh grade one decrescendo systolic murmur	He has a second in the decrescendo is heart murmur
"There is a herniation around the umbilicus"	There is a hernia under a umbilicus	There is lump at a right in a abdominal left
"Right upper quadrant tenderness"	Around a upper quadrant tenderness is	There is a right upper quadrant of a tenderness

transcriptions with audio recordings. Thirteen percent (52) of the 400 transcriptions contained errors. Single-word misspellings accounted for 21 (40%) of the transcription errors and single-word substitutions (e.g., "left" typed instead of "right") accounted for 12 (23%) of the errors. The remaining 19 errors (37%) were caused by the omission of words in the transcription. For instance, the experimenter omitted the word "ejection" when he transcribed the utterance "systolic ejection murmur in the right intercostal space."

To see whether the transcription errors distorted the output of the concept matcher, we observed whether the true utterance would have generated a different QMR list than did the transcribed utterance. Surprisingly, only one mistranscription caused a change in content of the QMR list. The most likely explanation for this low true error rate is that most transcription errors occurred with nonkeywords. Since the concept matcher performs keyword matching, it was unaffected by nonkeyword errors.

Speech-recognition Accuracy

We computed the overall recognition accuracy supported by each grammar by averaging the recognition accuracy for each of 400 utterances. We compared the overall accuracy of the two grammars and applied the paired t-test to gauge statistical significance. As the second column of Table 3 demonstrates, the recognition accuracy of the manually generated grammars was superior to the recognition accuracy of the programmatically generated grammars ($p < 0.0001$). Table 4 shows how a sample of input utterances was misrecognized by the two grammars.

Accuracy of Speech Recognition and Concept Matching Together

We measured the semantic accuracy of the interface program by processing utterances through both the speech recognizer and the concept matcher. We observed whether the "gold standard" QMR term—the term that the subject had both chosen during the experiment and rated as at least a general match—appeared on the ranked list of terms generated by the speech recognizer and the concept matcher. As shown in the third column of Table 3, the interface program using speech recognition would have generated a list that included the "gold standard" QMR term more than 70% of the time. Manually generated grammars supported the matching process significantly better than did the programmatically generated grammars (χ^2, $p < 0.0001$). Concept matching performed on utterances processed by the speech recognizer using manually generated grammars achieved accuracy of 81%, while concept matching performed on transcribed utterances achieved accuracy of 89%.

One potential explanation for the different performances of the two grammars is that utterances processed with one grammar could have generated more QMR terms to choose from than utterances processed with the other grammar. This would increase the probability that the "gold standard" QMR term appeared on the list. To investigate this possibility, we measured the number of QMR findings generated by each utterance. The last column of Table 3 shows that the programmatically generated grammars produced lists of QMR terms that were 2.45 findings longer, on average, than the lists produced by the manually generated grammars. This difference was highly significant (paired t-test, $p < 0.0001$). Human transcription of the utterance produced a list length of 8.53 ± 0.334 findings, which was not significantly different from that of the manually generated grammars. When comparing the two grammars, the manually generated grammars supported better matching

of utterances to QMR terms yet generated shorter lists of terms.

Discussion

This evaluation experiment using the Wizard-of-Oz paradigm yielded several interesting results. Although recognition accuracy alone was poor, the addition of concept matching boosted the system's semantic accuracy to 81% when manually generated grammars were used. This compared favorably with the 89% accuracy of concept matching alone, which did not depend on speech recognition. The 8% difference between the accuracy of the concept matcher with and without speech recognition shows the penalty that resulted from using speech input as opposed to keyword input.

These results should be interpreted in context. First, because we found significant differences in the recognition accuracy of the two grammars, it is possible that more accurate grammars could be created and could further improve semantic accuracy. Second, simple errors made by the concept matcher lowered the semantic accuracy for the entire system. If the concept matcher had had the word "gynecomastia" in its dictionary and had had a richer set of synonym pointers, it could have produced an exact or a general match for 96% of the utterances. This improvement in concept-matching accuracy would likely improve the semantic accuracy of the entire system.

On the other hand, if this speech interface were used in a real-world setting, it would likely have semantic accuracy that would be lower than that found in our experimental setting. The reason is that we created stimulus material to elicit clinical findings that were represented in the QMR vocabulary.

Thus, if our stimulus material and interface program had performed perfectly, we could have achieved 100% semantic accuracy. However, in a real-world setting, clinicians may wish to enter findings that do not exist in the controlled vocabulary. Thus, the upper limits of semantic accuracy will be determined by the expressivity of the controlled vocabulary, not by the speech-interface program.

It is debatable how accurate a speech system should be before it can be adopted in clinical practice. If accuracy is paramount to the success of a medical application, then semantic accuracy will need to be higher than we were able to achieve. However, if suboptimal accuracy is balanced by desirable attributes of speech input such as familiarity, ease of use, and expressivity, then our results might be acceptable. In our evaluation, we did not measure formally the value that users placed on using speech as opposed to typing or to manipulating a pointing device because subjects did not actually use the speech recognizer in the real-time experiment. Informally, however, a majority of subjects said that if our speech interface were faster (not aware that a hidden experimenter was actually manipulating the interface), they would prefer to use a speech interface for entering medical terms. Future experiments should formally evaluate the subjective benefits of using speech input so that these benefits can be compared with any performance penalty produced by speech misrecognition.

Other observations from this experiment were that it was feasible to use the Wizard-of-Oz paradigm for evaluating a speech-input system and that this experimental design provided an excellent way to measure a system's components, as well as to elicit standards by which the performance of the components could be measured.

The feasibility of Wizard-of-Oz experiments for the design of speech interfaces has been demonstrated by others.[13–15] However, to our knowledge, the feasibility of this experimental technique for the evaluation of a speech interface has never been demonstrated. The main concern regarding feasibility was whether physician subjects would believe that they were interacting solely with a computer and would speak to our interface as they would speak to any computer program in the clinical setting. The warm-up task (speaking simple phrases into the speech recognizer and observing the near-instantaneous output of the recognizer) seemed to convince subjects that they were interacting directly with a computer.

Later, when a hidden experimenter replaced the speech recognizer as the decoder of utterances, subjects experienced a brief delay while an utterance was being transcribed, but did not seem to suspect the cause of the delay. Therefore, we believe that subjects' utterances were representative of phrases that physicians would speak to a speech interface in a clinical setting.

Beyond feasibility, we believe that the Wizard-of-Oz technique has advantages as an evaluation strategy for speech interfaces because this technique allows analysis of different system configurations from one experiment. This technique allowed us to perform three different experiments: 1) how utterances transcribed by an experimenter match with QMR terms, 2) how utterances run through a speech system with manually generated grammars match with QMR terms, and 3) how utterances run through a speech system with programmatically generated grammars match

with QMR terms. We were able to measure overall system accuracy, as well as the contribution of each component, by transcribing utterances in the real-time experiment and later processing them with the speech recognizer and different grammars. In addition, because we compared the configurations on the same input, we were able to eliminate some potential confounding variables and thus study the target variables more accurately.

Another advantage of the Wizard-of-Oz technique is that it allows the experimenter to establish "gold standards" and then use those standards to measure the accuracy of the system's components. In this study, for instance, the real-time experiment characterized how subjects would react when their utterances were transcribed by an experimenter. Later, utterances run through the two grammars could be judged against this standard. The benefit of establishing ideal performance in the real-time experiment is that ideal performance becomes the benchmark for measuring semantic accuracy. Once the ideal action is known, the semantic accuracy of other system configurations can be measured by comparison with that ideal action.

The Wizard-of-Oz technique is not without its drawbacks, however. Although the technique can simulate the running of several experiments during one evaluation study, the technique is not the same as running several experiments serially. For instance, we assumed that the QMR term that a subject chose during the real-time experiment would be the term he or she would have chosen if his or her utterance had been run through the speech recognizer during the real-time experiment. This assumption might be valid if the list of QMR terms generated by speech recognition was a subset of the list generated by perfect recognition. However, because of misrecognition, terms appeared on the list that did not appear during the real-time experiment. Therefore, we could only guess what term the subject would have chosen in this situation; we assumed that he or she would have picked the same term that he or she chose in the real-time experiment, but we have no way to verify this assumption.

Another potential problem with this evaluation paradigm is the reliance on real-time human transcriptions of utterances for establishing "gold standards" for semantic accuracy. If the transcriptions contain many errors, then input to the system components will be distorted just as speech recognition might distort the input. Fortunately, in our experiment, only one mistranscription affected the output of the system. However, more significant penalties are possible if this technique is used for other evaluation experiments. Although we acknowledge these weaknesses in the experimental design, we believe that the benefits of this design, particularly the ability to compare different configurations on the same input, outweigh the limitations.

We evaluated a speech interface to a decision support system using a technique that has been used for design of speech systems but not for evaluation of speech systems. Using this design, we found that speech misrecognition imposed a 8% penalty on the semantic accuracy of our system. In addition, manually generated grammars supported better recognition for this limited domain. We also found that the Wizard-of-Oz experimental design offered many advantages for the evaluation of this interface and we believe that this technique may be useful to future evaluations of speech-input systems.

The authors thank Kevin Johnson and Alex Poon for their help with the design of the evaluation experiment; Nora Sweeny for editing an earlier version of this article; Randy Miller and Greg Cooper for assisting with the speech research at Stanford; and Camdat Corporation for providing the synonym dictionaries that are used in QMR.

References ■

1. Anderson JG, Aydin CE, Jay SJ, eds. Evaluating Health Care Information Systems: Methods and Applications. Thousand Oaks, CA: Sage Publications, 1994.
2. Wyatt J, Spiegelhalter D. Evaluating medical expert systems: what to test and how? Med Inf (Lond). 1990;15(3):205–17.
3. Forsythe DE, Buchanan BG. Broadening our approach to evaluating medical information systems. Proc Annu Symp Comput Appl Med Care. 1991:8–12.
4. Kaplan B, Duchon D. Combining qualitative and quantitative methods in information systems research. MIS Q. 1988;12:571–86.
5. Shiffman S, Detmer WM, Lane CD, Fagan LM. A continuous-speech interface to a decision support system: I. Techniques to accommodate for misrecognized input. J Am Med Informatics Assoc. 1995;1:36–45.
6. Shiffman S, Lane CD, Johnson KB, Fagan LM. The integration of a continuous-speech-recognition system with the QMR diagnostic program. Proc Annu Symp Comput Appl Med Care. 1992:767–71.
7. Miller RA, McNeil MA, Challinor SM, Masarie F Jr, Myers JD. The INTERNIST-1/QUICK MEDICAL REFERENCE project—status report. West J Med. 1986;145(6):816–22.
8. Hauptmann AG, Rudnicky AI. Talking to computers: an empirical investigation. Int J Man-Mach Stud. 1988;28(6):583–604.
9. Rudnicky AI, Sakamoto M, Polifroni JH. Spoken language interaction is a goal-directed task. In: 1990 International Conference on Acoustics, Speech and Signal Processing. Albuquerque, NM: IEEE, 1990:45–8.
10. Kuhn K, Gaus W, Wechsler JG, et al. Structured reporting of medical findings: evaluation of a system in gastroenterology.

Methods Inf Med. 1992;31(4):268–74.
11. Massey BT, Geenen JE, Hogan WJ. Evaluation of a voice recognition system for generation of therapeutic ERCP reports. Gastrointest Endosc. 1991;37(6):617–20.
12. Johnson K, Poon A, Shiffman S, Lin R, Fagan L. A history-taking system that uses continuous speech recognition. Proc Annu Symp Comput Appl Med Care. 1992:757–61.
13. Dahlback N, Jonsson A, Ahrenberg L. Wizard of Oz studies—why and how. Knowl-based Syst. 1993;6(4):258–66.
14. Gould FD, Conti J, Hovanyecz T. Composing letters with a simulated listening typewriter. Communications of the ACM. 1981;26:295–308.
15. Isaacs E, Wulfman CE, Rohn JA, Lane CD, Fagan LM. Graphical access to medical expert systems: IV. Experiments to determine the role of spoken input. Methods Inf Med. 1993; 32(1):18–32.
16. Coutaz J, Salber D, Balbo S. Towards automatic evaluation of multimodal user interfaces. Knowl-based Syst. 1993;6(4):267–74.
17. Wyatt JC, Detmer WM, Fagan LM. Design and evaluation of multimedia stimuli to evoke clinical concepts. Proc Annu Symp Comput Appl Med Care. 1993:834–8.
18. Lee K. Automatic Speech Recognition: The Development of the SPHINX System. Boston: Kluwer Academic Publishers, 1989.
19. Hooper RS. Indexer Consistency Tests: Origin, Measurement, Results, and Utilization. Technical report 95-56. Bethesda, MD: IBM Corporation, 1965.