Opportunities to Use Speech Recognition for Bridge Inspection

Jirapon Sunkpho¹, James H. Garrett, Jr.,², Member, ASCE, and Asim Smailagic³

Abstract

Inspections of bridge mandated by the National Bridge Inventory (NBI) are required to be performed every two years. On an NBI inspection, inspectors go out to the field and collect inspection information based on the requirement of the state owning the bridge. Since the development of mobile computing devices, such as notebook computers, personal digital assistants (PDAs), and pen computers, many organizations have been trying to deliver computing support for bridge inspectors in the field using these devices in order to improve productivity. However, some bridge inspectors are still using paper-based forms and clipboards during their inspection activity. From our experience with bridge inspectors from the Pennsylvania Department of Transportation, District 11, the user interface modality as well as the mobility of the hardware have major effects on the usability of such systems. It has been long known that the success or failure of a particular application depends significantly on the way the user interacts with the system. We have studied the use of a wearable computer to provide an unobtrusive hardware platform supporting bridge inspection. This paper specifically discusses the potential use of speech recognition for the bridge inspection application in order to improve the usability of the user interface. The background of speech recognition technology, along with the results of our preliminary study will be discussed in this paper.

Key Words

Bridge inspection, speech recognition, field computing, wearable computer

¹ Doctoral Student. Dept. of Civil and Environmental Engineering, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA, 15213
² Prof., Dept. of Civil and Environmental Engineering, Institute of Complex Engineered System, Carnegie Mellon University
³ Research Engineer, Institute of Complex Engineered System, Carnegie Mellon University
Introduction

Bridges are one of the major types of infrastructure requiring a significant amount of inspection. There are nearly 590,000 bridges over 6.1 m in length in the National Bridge Inventory (NBI) that must be inspected at least every two years. More than a fourth of all bridges are more than 50 years of age and one out of three bridges in America is rated structurally deficient or functionally obsolete [9]. Bridges have to be inspected in order to ensure that those bridges meet the functional and safety requirements placed upon them. Inspection is a data collection process and is one of the major process components in any infrastructure management system. Data collected are used to support maintenance, redesign and replacement decisions.

Many states, research organizations, and commercial companies have addressed this issue by providing bridge inspectors with computing devices such as personal digital assistants (PDAs), notebooks and wearable computers [5, 7, 13]. However, the computing support needed and usable in the field is different from that needed and usable in the office environment. First, the hardware is not a regular desktop computer where the user interacts with the computer via keyboard and mouse. Most of the computers that are currently being used in the field, especially in bridge inspection, are pen-based systems where the user interacts with the computer via pen and touch-screen interface. These computer systems mimic the inspection forms and require the inspectors to perform the data entry via the use of a pen interface and handwriting recognition software. There are some drawbacks associated with using a pen-based interface for PDAs and wearable computer [6]. First of all, the size of the screen on a PDA and wearable computer is small. The pen-based interface requires enough space on the display for the user to write, which puts a limit on the other controls that can be placed on the Graphical User Interface (GUI). Secondly, handwriting is a slow way to input data and prone to error. Thus, the use of a pen-based interface may not reduce the data entry time. Another drawback of using a pen-based interface is that it requires the use of both hands to operate it. Many inspection activities require hands-free or near hands-free operation for both safety and efficiency reasons, for instance, when an inspector is measuring the size of a defect on a bridge.

Speech technology, both recognition and synthesis, is an emerging technology that allows computer users to communicate with the computer more naturally. Reported advantages of speech interfaces over other interaction methods, such as a keyboard, hand writing, or a pointing device, include the naturalness of interaction, the speed of the text produced, and the elimination of the use of hands [4, 11, 12]. Speech recognition systems have also been found to be advantageous in situations where eyes and hands are occupied with control and monitoring functions and when only a limited keyboard and/or screen is available [3]. Speech is considered more expressive and efficient than text, as it places less cognitive demands on the speaker and permits more attention to be devoted to the context of the message [2]. In bridge inspection, where the data collected is the most important information supporting the decisions about maintenance, redesign and replacement, the expressiveness of the information plays a critical role in determining the quality of the data collected. In addition to a photo, sketch, or other piece of multimedia
information, a textual comment about the condition of a specific bridge element provides another dimension to the collected data. This textual comment produced by speech can be more expressive and thus better describe the condition of the bridge elements.

This paper will discuss the possibility of using speech recognition technology in bridge inspection. The issues regarding how to apply speech recognition technology to a particular application will also be discussed. A preliminary study of the use of speech recognition in a data entry task in bridge inspection will be discussed in the last section. The following section will provide a description of the speech recognition technology terms and definitions.

**Speech Recognition Terms and Definitions**

Speech recognition system characteristics are discussed elsewhere in the literature [3, 14, 15, 16, 18]. The following section contains the important characteristics in speech recognition technology: speech flow, speaker model, vocabulary size, enrollment and training, feedback, and error correction methods.

**Speech Flow**

This parameter represents the flow of the speech from the user speaking to the speech recognition system. Earlier speech recognition systems using *discrete speech recognition* require the user to pause briefly for approximately 100 – 200 milliseconds between each word. Speech recognizers capable of accepting *connected speech* input appeared on the market in the early 1980s. Connected word systems are able to recognize words without the speaker having to pause between words, however, the individual words must be spoken with the same intonation pattern that would be used if they were read from a list. Most of the speech recognition systems today use *continuous speech recognition*, which allows users to speak without pausing between words and use natural speech rhythm and intonation.

**Speaker Model**

The speaker model refers to the extent to which the system must have the data about the voice characteristics of the particular human speaker who is using the system. The *speaker-dependent* system requires specific examples of a user’s speech pattern. Such systems must have a sample of how a user speaks before the system can recognize the speech from that speaker. Thus, for such systems, a user must train, or enroll to, the system before using it. *Speaker-independent* systems do not require examples of speech from the user. However, to achieve the certain level of accuracy, the size of the vocabulary has to be small enough and each word has to be acoustically distinct. This system is useful when the user population is large and the number of uses is infrequent, e.g. application for the telephone company. *Speaker adaptive* systems begin as speaker-independent systems but are able to adapt to users as they use and correct the errors that are produced by the system.
Vocabulary Size

The vocabulary size is simply the number of words that the system can recognize. Current commercial speech recognition systems, such as IBM ViaVoice™ or Dragon Naturally Speaking™, contain a base vocabulary of as many as 260,000 words and it can be made to recognize words in an active vocabulary of as many as 64,000 words. An active vocabulary is a set of words that the recognizer has to choose from at anytime during the recognition process, while the base vocabulary is the total number of words that are stored in the recognition system’s database. The size of the vocabulary can greatly affect the performance of the speech recognition system because the larger the vocabulary, the more likely that the system will get confused trying to identify what word the user is speaking.

Enrollment and Training

Enrollment is the process of providing speech templates to the recognition system for the different words in the vocabulary [18]. The enrollment process is required by the speaker-dependent systems. For the speaker-independent systems, the vendor who developed the speech recognition engine provides the templates that are believed to achieve the highest accuracy for a large population of potential users. Speaker adaptive systems start out with the vendor-enrolled template, allowing the user of the system to talk to the system immediately. However, speaker adaptive systems are capable of performing enrollment as well, in order to improve the accuracy.

Feedback

Because speech recognition systems make errors either due to the technology or the use of the system, the user must be provided with some type of feedback by which errors can be detected [17]. Two levels of feedback and two modalities in which feedback can be presented are described in literature. The two levels of feedback are termed primary and secondary [1, 10, 17]. Primary feedback occurs when the system responds immediately to a recognized command or vocabulary item. For example, when a bridge inspector says, “insert photo,” the system immediately responds by downloading a photo from the camera. Secondary feedback occurs when the command or vocabulary items are repeated back to the user for verification before any action is taken by the system. For example, before a photograph will be downloaded from the camera, the system repeats the photo number and waits for confirmation from the inspector. The modalities in which feedback can be presented are via the auditory channel, the visual channel, or both. The best modality for providing feedback depends on the environment and tasks being performed. Inappropriate feedback may interfere with the primary task and thus can degrade the performance.

Error Correction

Once the user of the speech recognition system detects errors, the system has to allow the user to recover from those errors. Four types of error correction methods were described...
in [10]: immediate reentry; backtrack and delete; backtrack and selective editing; and cancellation and re-entry. **Immediate reentry** allows the user to repeat the incorrect words after the user detects an error via visual feedback. **Backtrack and delete** allows the terminal item of the continuous string to be deleted. **Backtrack and selective editing** allows the user to edit a misrecognized phrase by positioning a cursor over an incorrect entry and entering the correct term. **Cancellation and re-entry** allows the user to reject the recognized phrase using a single command and then re-enter the phrase. The system can also give the list of possible candidate words and allow the user to choose from that list.

*Opportunities to Use Speech Recognition for Bridge Inspection in the Field*

A speech recognition system can address two keys concerns in human-computer interaction: the demands for ease of use and constraints on the user’s ability to work with a keyboard or other pointing devices [8]. In bridge inspection, the possible uses of speech recognition are: free form data entry; selection from a list, speech activated commands; integration with other interface modalities; and hand-busy data collection.

**Free Form Data Entry**

One of the most important tasks in bridge inspection is to produce the inspection report for filing and documentation. Bridge inspectors need to fill out the inspection form in order to complete an inspection for a particular bridge. The information that will get into the form includes the condition rating of a particular element or a narrative description of the condition of a particular bridge element. The narrative description is a free-form data entry type, which means there is no specific format for entering the information. Bridge inspectors describe the condition of a bridge element as they see. Speech is an excellent method for composing strings of text due to the speed and the expressiveness of speech [4]. With a reasonable accuracy rate, speech can be a very promising way to perform a free-form data entry.

**Selection from a List of Phrases**

Many times, to enter the inspection information, bridge inspectors have to select appropriate information from lists of possible entries. For example, to specify the condition rating of a bridge element, bridge inspectors pick a number from 1 to 9. As also pointed out in [4], speech is an excellent way of performing 1 out of N selection. Speech will be especially effective when N is on the order of hundreds since all possible entries could not be displayed at the same time in a graphical user interface and would require the user to scroll up and down.

**Speech Activated Commands**

The simplest form of speech activated commands are the commands that map exactly onto the GUI. In other words, what the user sees in the GUI can be done simply by issuing a speech command by the same name as the visual command icon. By doing this, the user of the speech recognition system has to memorize the set of commands that can
be spoken. The more complex form of the commands is the natural language commands, where all possible ways of issuing commands by the user have been programmed to be speech recognized. The natural language commands make the program more friendly and put less cognitive load on the bridge inspectors.

**Augment other Interface Modalities**

In bridge inspection tasks where the coordinates or the location of an inspected object needs to be specified, speech can be used to augment other types of interaction supported by the system, such as a pen interface. The user can use a pen or stylus to indicate the location of the inspected element and use speech to provide a detailed description of the inspected bridge element.

**Hand-busy/Eye-busy Data Collection Tasks**

Due to the nature of inspection, the inspectors need to have their hands free to support their climbing and inspecting activities [7]. A speech interface provides an unobtrusive mode of system operation. Users of a speech interface can interact with the computer while their hands and eyes are busy. Bridge inspectors also wear gloves while inspecting the bridge, thus it is not easy for them to interact with the computer using a pen-based or keyboard interface. Thus, providing hands-free and unobtrusive operation to bridge inspectors in the field will likely improve their job performance and inspection productivity.

**Integrating Speech Recognition in Bridge Inspection**

Although speech is an attractive technology that allows a user of a software system to interact with the system in a more natural and hands-free manner, speech technology is not a technology that one can buy “off the shelf”, simply integrate into a system, and obtain improvement in overall efficiency of the system. In integrating speech recognition technology, several decisions have to be made: what speech recognition device will be used; what kind of feedback and error correction methods will be provided; what type of speech flow and speaker model should be used; how the vocabulary will be designed; what types of tasks should be performed using speech recognition, etc. In order to answer these questions, an understanding of how limitations and abilities of users interact with the limitations and functions of the recognition system in the context of a particular physical environment are essential [15]. Unfortunately, there is no general rule of thumb that can be employed when one wants to apply speech recognition technology to a particular application. As mentioned earlier, the four main factors that govern the use of this technology are: the speech technology itself; humans who operate the system; the application; and the physical environment. Speech recognition technology itself has some limitations. Talking to the computer is not the same as talking to another human. A speech recognition system relies on the speech signal produced by the speaker and transforms that signal into the recognized text by using some sort of pattern recognition algorithm [15]. Thus, when the signal produced by the speaker does not match the signal in the system database, different kinds of error will occur. Therefore, one should not use
a different microphone in the field from that used when enrolling the speech recognition system, because different microphone may produce different sounds that might cause recognition errors. The effectiveness of any recognition system depends on the human who provides the speech input and makes use of the technology. Thus, it is clear that human factors will play an important role in the design of any speech interface. Knowledge and experience about the recognition system, cognitive demands, effort, stress, and fatigue can clearly affect the performance that may be achieved with the recognition system [10, 11, 15]. The effect of the physical environment plays an important role on how a human operator produces speech and might also distort the speech signal [11, 15]. In bridge inspection application, traffic noise is one of the important factors that needs to be considered. Another important factor that affects the performance of the speech recognition system is the application environment. The application environment consists of the tasks and application for which speech recognition is being used. It is very important to choose appropriate tasks and applications for speech recognition. For example, it is well known that speech is not a good interface for drawing an object. Therefore, in bridge inspection, one might not want the bridge inspectors to use speech to create a sketch.

From our experience with bridge inspectors from the Pennsylvania Department of Transportation District 11, we learned that in order to improve inspection productivity, they need to be able to complete an inspection form before they leave the inspection site. Since, speech provides an excellent means by which compose text strings, therefore, we have integrated the ability to create narrative comments about a bridge element using speech. Also, in order to provide hands-free operation, we allow them to navigate to different parts of the inspection form using speech. In the next section, we will describe our preliminary study of using speech recognition to perform these tasks.

Preliminary Work

In order to demonstrate the potential uses of speech recognition technology for bridge inspection, a prototype application for bridge inspection has been developed using Visual Basic™. Speech recognition was integrated into the prototype. The speech recognition engine used is IBM ViaVoice™. The system has a large vocabulary (64,000 words), continuous speech recognition, and is speaker dependent. Feedback from the system is visually provided. Three types of error correction methods are implemented: immediate reentry, backtrack and delete, and backtrack and selective editing. Figure 1 shows the interface of the prototype used during the preliminary study. At the top of the interface, visual feedback was provided for the recognized commands. When the user dictates the inspection comment, visual display of the recognized texts are shown in the upper text box. Bridge elements that need to be inspected are shown in the two list boxes on the left. The lower text box shows the narrative comments about the inspected element from last inspection conducted.

During this preliminary study, three volunteers from the Department of Civil Engineering were asked to fill out the inspection form, which is one of the tasks that needs to be performed in the field during inspection. The subjects performed the task using both a
speech interface and a pen interface. The time to complete the task along with the user satisfaction rating were measured. Speech recognition errors and task errors were also recorded. The user satisfaction rating scale is shown in Figure 2.

Prior to the test, the subjects had to train the speech recognition system in order for the recognition system to recognize their voices. The training session took approximately 30 minutes. After the speech recognition system training, the subjects were introduced as to how to perform the task both using a speech interface and a pen interface.

During the test, each subject was given a copy of a previous inspection report. The task they were asked to perform was to fill out the information about the abutment given in the inspection report. Two of the three subjects performed the task using the speech interface first and then the pen interface. The third subject performed the task using the pen interface first and then the speech interface. When the subjects performed the task using speech interface, the subjects also navigated to different parts of the form using speech by saying “go to <name of the element>.” The subjects performed the task using the pen interface by either using handwriting recognition or the virtual keyboard. To navigate to different bridge elements, the subjects simply clicked on that particular element.

The results of this preliminary study are shown in Table 1. Note that the results shown in Table 1 are the average results from all the three subjects; individual results are shown in italic for all three subjects. From Table 1, we can see that the time needed for data entry was significantly less when using speech (about 2 times faster). The user satisfaction rating of the speech recognition interface was also higher than that of the pen interface. However, the error for the speech interface was a little bit higher than that for the pen-based system. This is due to the fact that the speech recognition engine could not achieve...
100% recognition accuracy. Also, the subjects did not correct the errors resulting from the speech recognition engine due to either not noticing the errors or the error correction methods not working very well for them. The average accuracy of the speech recognition system during the experiment was 92%.

<table>
<thead>
<tr>
<th></th>
<th>Time (minutes)</th>
<th>User Satisfaction Rating</th>
<th>Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech Interface</td>
<td>3:26</td>
<td>4.56</td>
<td>4.87</td>
</tr>
<tr>
<td></td>
<td>(3:25, 4:23, 2:30)</td>
<td>(4.53, 3.15, 6)</td>
<td>(6.09, 4.87, 3.65)</td>
</tr>
<tr>
<td>Pen Interface</td>
<td>7:54</td>
<td>3.54</td>
<td>1.23</td>
</tr>
<tr>
<td></td>
<td>(7:00, 5:36, 11:05)</td>
<td>(3.23, 3.38, 4.00)</td>
<td>(2.47, 1.22, 0)</td>
</tr>
</tbody>
</table>

Table 1. The Results from the Preliminary Study

Conclusions

One of the problems of providing computing support for field inspectors today is the usability of the user interface. This paper discussed the problem of the most commonly used interaction mechanism for supporting bridge inspectors in the field, the pen-based interface. Speech recognition technology can overcome some of the drawbacks of the pen-based interaction mechanism, such as the lack of scalability, low speed and low accuracy.

Speech recognition terms and definitions were introduced. This includes speech flow, speaker model, vocabulary size, enrollment and training, feedback, and error correction methods. The possible uses of speech recognition in bridge inspection were also discussed. In applying speech recognition technology, four main factors are needed to be understood: speech recognition technology; human factors; the physical environment; and the application environment. An understanding of these factors will help determine the design of an efficient speech recognition system.

Finally, a small preliminary experiment has been performed to study the use of speech recognition in data entry. The results are very promising in terms of speech and user satisfaction. However, the experiment was performed for only few subjects and in a controlled environment, where the real conditions of the field environment might degrade the use of speech recognition. Therefore, the need for evaluating the applicability of speech recognition in the field with a larger number of real bridge inspectors is mandatory. Such a study will also help us to understand the physical environment as well as the human factor aspects of this application.

Acknowledgements

The authors would like to thank Mr. Marty Neaman from the Pennsylvania Department of Transportation District 11 office and Mr. Pat Kane from the Michael Baker Corporation for their input in to the development of this prototype. The authors would also like to thank Heather MacLean, Grant Bromhal, and Rebecca Bucheit for their participation during the preliminary study. A stay at Carnegie Mellon University for
Jirapon Sunkpho are sponsored by the Royal Thai Government. Funding for this work was provided through the Pennsylvania Infrastructure Technology Alliance.

References