SPEECH INTERFACES FOR MOBILE COMMUNICATIONS

Hirotaka Nakano

NTT DoCoMo Multimedia Laboratories
3-5 Hikarinooka, Yokosuka, Kanagawa, 239-8536 Japan
E-mail:nakano@mml.yrp.nttdocomo.co.jp

ABSTRACT

This paper will explain speech interfaces for mobile communication. Mobile interfaces have three important design rules. Do not disturb the user's main task, work within the restrictions of user's ability, and minimize the resource requirements. Social acceptance is also important. In Japan, trial and regular services with speech interfaces in mobile environments have already been launched, but they are not widely used. They must be improved in mobile interfaces. The speech interface will not replace web browsers, but should support and interwork with other interfaces. We also have to discover contents that suit speech interfaces.

1. INTRODUCTION

This paper presents a voice interface for mobile environments, especially cellular phones.

2. MOBILE INTERFACE

There are three important goals in designing a mobile interface.

First, it must not disturb the user's main task (walking, driving, etc.). It must be easy to use, and must support interrupts to handle higher priority tasks.

Second, it must work within the restrictions imposed by the user's ability and his environment as well as changes in both of these characteristics. For example, a user may not be able to read the regular display because of car movement, he may not be able to send a voice message because of loud background noise or may not be able to type because his cold fingers are too stiff. (Fig 1.)

Third, it must minimize the requirements placed on the user's resources such as CPU power, display quality, battery lifetime, and radio interface performance.

3. SOCIAL ACCEPTANCE

Social acceptance is also important. First, the user must pay attention to his surroundings; his usage of the interface must not trouble others and must not look strange. Privacy is also a big concern. Personal information should be processed in a secure manner.

We note that cellular phone usage is restricted in hospitals and crowded trains because of the risk posed to medical equipment and pacemakers.

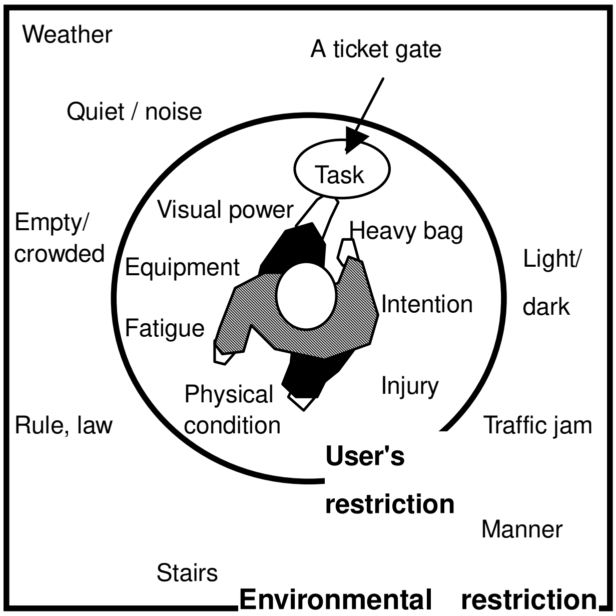

Fig. 1. Mobile environment.

4. RECENT SERVICE

In Japan, trial and regular services with speech interfaces have already been launched.

A voice dial and voice command system has been embedded in a cellular phone. There is a handy mail terminal that can read e-mail aloud, but there are not many users.

We also have a simple voice information service that uses isolated word recognition for sightseeing (input; a place of interest, output; a guide), weather (input; location, output; weather forecast), transfer information (input; departure and destination station, output; necessary time, route and the fare). This also sees little demand.

A ring tone melody download service is somewhat popular, and is a good use of ASR. In this case, the user can be expected to know the desired song's name exactly, so ASR is more easier to use than selecting from a huge list on a display; the recognition rate is important.

Several voice portal trials involving cellular phones are currently underway. For instance, "Voizi" (Japan Telecom Co., Ltd) and "V portal" (NTT communications Corp.) are standard voice portal services, and "ITS voice portal" (KDDI Corp.) is a specialized navigation service that uses cellular phones with built-in GPS. Their success remains unclear.

5. PROBLEMS AND SOLUTIONS

As noted above, speech interfaces in mobile environments are not widely used in Japan. There are several main reasons.

5.1. Few chance

The speech interface seems useful, but in real daily life we have few chances to use it because we hesitate to make noise in a public space and to disclose private information. By comparison, text mail on a cellular phone doesn't make noise to read or write messages, and user's privacy is usually maintained. As a result, cellular phone e-mail is popular in densely populated Japanese cities.

There is also the cultural aspect, Japanese prefer text mail to voice mail. Voice mail dose not come into wide use in office. Before the mobile web-phone appear, a public voice message service was popular for a period among young people in Japan, it was used to make communities. Now users have switched to mobile web-phones with text.

5.2. Communication difficulty

Current speech interfaces offer insufficient recognition rates because of changeable background noise, limited voice quality affected from coding and transmission error, and the weak CPU power and memory resources if processing is done locally.

Poor browsing and editing performance prevents the interfaces from becoming popular. It is hard to allocate editing points, make an edit, and then check it [1]. Users usually must listen to voice responses attentively because it is difficult to have the response repeated. On the other hand, rereading a text is easy even if the user in a noisy place.

It is also hard to list many items through a speech interface. This makes it difficult to achieve a sophisticated information service because the user has to memorize many items.

5.3. Higher load

Speech interfaces are excellent in terms of their low physical interference. But mental interference cannot be ignored. We have to pay attention if we are to speak a message accurately. Typing offers better accuracy. Speech interfaces suit transcription, not composition. Key operation coexists with user's thinking, but voice commands interrupt it [2].

The speech information service has a stiff dialogue, so the user cannot take the initiative. The telephone-based voice information service usually cannot well handle interruptions unlike Web browser operations.

Speech interfaces demand that the ASR system be able to distinguish the user's voice input from other voices and noise. The user has to tell the system which statement is for ASR, and which is not.

Speech is not good for making copies. Voice memos occupy a lot of resources.

Many different Japanese words have the same sound and with speech it is difficult to clearly differentiate them.

5.4. Cost

In telecommunication services, most speech recognition engines are run on network servers. Accordingly, the customer is sensitive to the call charge, particularly with cellular phones. The recognition sever is expensive if many telephone circuits must be handled during peak loads. If the engine is run on the cellular phone, the terminal cost is high.

5.5. Solutions

Given these problems, higher recognition rates are needed, and the microphone must be set near user's mouth to minimize the social problem and improve recognition performance. The speech interface should allow interruptions, and activation of another interface. The interface should also display the recognized text rather than use just voice response. More flexible dialogue recognition is needed.

Several parts of the cost problem will be improved by introducing Voice over IP (VoIP). With VoIP, we can handle voice data the same as other digital data. A cellular phone can use the packet network to send a voice command, so users don't have to make an effort to minimize the connection time. The server can use ordinary network devices, and doesn't have to support many telephone circuits.

6. THE INTRODUCTION OF SPEECH INTERFACES

The introduction of speech interfaces into cellular phones will proceed in steps. The first step is voice dialing and voice commands. These functions are already in use, but improved performance is needed. The next step is to assist in text input. The user makes a draft with ASR, and then edits the text via the keys. The final stage is complete replacement of the keyboard. This needs an intermediate processor that can catch the story from the dialogue and then compose the appropriate message.

7. CONCLUSION

The basic concept of our mobile web-phone "i-mode" is " my concierge ", which is very suitable for speech interface usage. But this interface has certain constraints on usage as mentioned above. VoiceXML can be used in automated telephone systems, but it will not replace web browsers. Speech interface will have to support web browsing and cooperate with other interfaces. We also have to discover contents that suit the speech interface.

One possibility is to allow access to current popular services through a speech interface. For example, the most popular service, Peer-to-Peer (P2P) communication (telephone calls and e-mail) can be assisted by speech interface to improve ease of use. Depending on the user's situation, a suitable interface is selected for making a call, creating, and sending an e-mail. A community service is supported in a same way with P2P communication. Game services and software pets sometime don't need high recognition rates (most users will accept a pet that doesn't follow every command), and so are good speech interface applications.

Seniors, who are expected to be the new breed of cellular phone users, will have strong demand for an effective speech interface [3]. Cellular phones for senior users are being market by several carriers, and sales are good.

[1] C. Karat, C. Halverson, D. Horn, and J. Karat, Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition Systems, In *Proceeding of CHI'99: Human Factors in Computing Systems* (Pittsburgh, May 15-20). ACM Press, New York, 1999, 568-575.

[2] B. Shneiderman, The limits of Speech Recognition, *Communications of the ACM, 43, 9* (Sep. 2000) 63-65

[3] http://www.jmm.com/xp/jmm/press/2001/pr_090601.xml