Speech Interface Exploiting Intentionally-Controlled Nonverbal Speech Information

Masataka Goto
National Institute of Advanced Industrial Science and Technology (AIST).
Ibaraki 305-8568, Japan
m.goto@aist.go.jp

Katunobu Itou
Nagoya University.
Aichi 464-8603, Japan
itou@is.nagoya-u.ac.jp

Tetsunori Kobayashi
Waseda University.
Tokyo 169-8555, Japan
koba@waseda.jp

ABSTRACT
This paper describes our research on speech interfaces using nonverbal speech information. Although speech information consists of verbal and nonverbal information, most speech-recognition research has made use of only verbal information such as words and sentences. From among nonverbal information, we have focused on hesitation (filled pause) and prosody (voice pitch) to create four speech-interface functions: *Speech Completion*, *Speech Shift*, *Speech Starter*, and *Speech Spotter*. These functions can be used without the need for training because they are based on simple, easy-to-utter rules on how to utter nonverbal information to invoke them. By having users intentionally utter nonverbal information according to those rules, we have achieved interfaces that can exploit the potential of speech in various forms.

ACM Classification: H5.2 [Information interfaces and presentation]: User Interfaces. - Voice I/O.
General Terms: Design, Human Factors
Keywords: Speech interface, speech recognition, nonverbal information, filled pause, voice pitch

1 INTRODUCTION

Most speech recognition research has focused on ways of obtaining verbal information such as phonemes and words from speech and on improving the speech-recognition rate. The technology developed for this purpose is, of course, important, but raising the recognition rate by itself is insufficient — interfaces using speech recognition are still difficult to use. With the goal of using speech recognition through a comfortable and easy-to-use interface, we set out to achieve speech-interface functions that can tap the potential of speech by making full use of *nonverbal information*, which has mostly been ignored by speech recognizers.

Nonverbal information such as hesitation and voice pitch has, if anything, been considered a problem in that it can cause recognition errors. For example, erroneous recognition and inappropriate input caused by hesitation during speech input is a common occurrence. Although some attempts have been made at using voice pitch (or prosody), mainly to improve speech recognition rates [1, 2] or to analyze phrase boundaries [3], these attempts dealt with auxiliary aspects of nonverbal information unintentionally contained in natural speech input. Although there were a few exceptions [4, 5] that used nonverbal information for interface functions, they did not exploit both verbal and nonverbal information.

We developed the four speech-recognition-based speech-interface functions shown in Figure 1 by using two kinds of nonverbal information, a *filled pause (vowel-lengthening hesitation)*[1] and *voice pitch (fundamental frequency, F0)*, both of which are unique to speech, having no counterparts in writing. In contrast to conventional speech input, which conveys only verbal information, our functions place importance on intentional hesitation or pitch changing by the user during speech input. This has made it possible to convey new types of information from the user side to the computer side and to achieve interface functions that exploit features unique to speech.

2 Speech Completion: On-demand completion assistance using filled pauses

Speech completion is a function that helps a user enter a word or phrase by *completing* (filling in the rest of) a phrase fragment uttered by the user, as depicted in Figure 1. Although the concept of completion is widely used in text-based interfaces [7], there have been no reports of completion being effectively applied to speech. By using a filled pause, we enable a user to effortlessly invoke the speech-completion function, which helps the user recall uncertain (difficult to remember) phrases and saves effort when the input phrase is long. When a user hesitates by lengthening a vowel sound (by uttering a filled pause) during a phrase (like the Japanese phrase "*maikeru–*"[2] corresponding to "Michael, uh..." or "Michael–" in English), our system immediately displays completion candidates whose beginnings acoustically resemble the uttered fragment (like "Michael Jackson", "Michael McDonald", etc.) so that the user can select the correct one. Experiments with 45 subjects on a system consisting of a filled-pause detector and an extended speech recognizer capable of listing candidates showed the effectiveness of this function.

3 Speech Shift: Direct speech-input-mode switching through intentional control of voice pitch

Speech shift is a function that enables a user to enter the same word with it having different meanings (functions) by inten-

[1] Although there are several hesitation phenomena [6], we target only filled pauses, which we here use in the meaning of prolongation of a vowel sound, like "er..." (underlining indicates a filled pause).

[2] When a foreign name like "Michael Jackson" is pronounced in Japanese, it is regularized to conform to the Japanese style: "*maikeru jakuson*". In Japanese, vowel-lengthening hesitations like "maikeru–" are common.

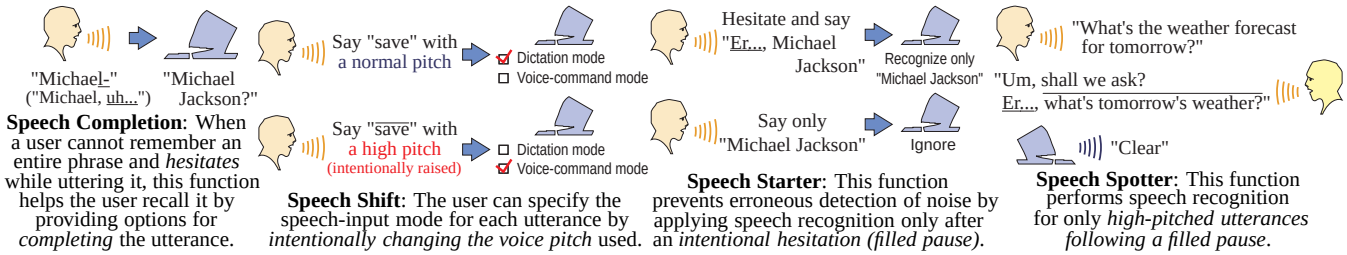

Figure 1: Four speech-interface functions exploiting nonverbal speech information. Underlining indicates a filled pause (vowel-lengthening hesitation) and overlining indicates that the pitch is intentionally raised.

tionally controlling the voice pitch used. Current speech-input interfaces cannot distinguish a word from the same one with a different pitch because they recognize only verbal information. As depicted in Figure 1, the speech-shift function can distinguish an utterance with a high (shifted) pitch from one with a normal (low) pitch and assign them to different speech-input modes. On a voice-enabled word processor, for example, it can regard the former type utterance as *voice-command-mode input* (such as file-menu and edit-menu commands like "delete") and the latter type as *regular dictation-mode text input*. Experiments with 20 subjects showed that the speech-shift function is an effective, easy-to-use, and labor-saving input method.

4 Speech Starter: Noise-robust endpoint detection using filled pauses

Speech starter is a function that enables noise-robust endpoint (utterance) detection for speech recognition in non-stationary noisy environments. When current speech recognizers are used in a noisy environment, a typical recognition error is caused by incorrect endpoints because their automatic detection is likely to be disturbed by non-stationary noise. As depicted in Figure 1, the speech-starter function enables a user to specify the beginning of each utterance with an intentional hesitation (a filled pause), which is used as a trigger to start speech recognition. Since filled pauses can be detected robustly in a noisy environment, practical, hands-free endpoint detection without using any input device other than a microphone can be achieved. Experiments showed that this function provided good performances in noisy conditions at SNRs of 0 and 10 dB.

5 Speech Spotter: On-demand speech recognition in human-to-human conversation

Speech spotter is a function that enables a user to enter voice commands into a speech recognizer during the course of a natural human-to-human conversation. In the past, it has been difficult to use automatic speech recognition in human-to-human conversation situations because it was difficult to judge, from microphone input only,[3] whether the user was speaking to another person or the speech recognizer. We solve this problem by enabling a user to *intentionally* indicate whether each utterance is to be accepted (processed) by the speech recognizer by using two kinds of nonverbal information, a filled pause and voice pitch. As depicted in Figure 1, the speech-spotter function regards a user utterance as a command utterance only when it is uttered with a high pitch just after a filled pause (like "er..."). Using this function, we have built two application systems: an on-demand information system for assisting human-to-human conversation and a music-playback system for enriching telephone conversation. Experiments showed that the speech-spotter function is robust and convenient enough to be used in both face-to-face and cellular-phone conversations.

[3] Our speech-spotter function can be considered a hands-free version of *dual-purpose speech* [8], which uses a push-to-talk button.

6 CONCLUSION

We have described four novel, easy-to-use speech interfaces that have users intentionally utter nonverbal information. We believe that the ability of speech to simultaneously convey both verbal and nonverbal information is of fundamental importance, and have created four speech-interface functions that use both of these to full effect. We found that it is possible to use nonverbal information to achieve new interface functions by having users utter that information intentionally. This differs with past approaches of using nonverbal information in which the information is unconsciously uttered as simply a supplement to verbal information.

Our experiments showed that all four functions are robust and effective for Japanese speech. In the future, we will apply these functions to other languages because the underlying ideas of these functions are universal and language-independent. In addition, we plan to study the synergies between these four separate functions and build integrated interfaces. Future work will also include further development of our research toward diverse speech-interface functions that exploit the great potential of speech.

ACKNOWLEDGMENTS

The authors would like to thank Koji Kitayama, Yukihiro Omoto, Satoru Hayamizu, and Tomoyosi Akiba for their valuable discussions and collaborative research on speech interfaces.

REFERENCES

1. A. Waibel. Prosodic knowledge sources for word hypothesization in a continuous speech recognition system. *Proc. of ICASSP 87*, pp. 856–859, 1987.
2. A. Stolcke, E. Shriberg, D. Hakkani-Tür, and G. Tür. Modeling the prosody of hidden events for improved word recognition. *Proc. of Eurospeech '99*, pp. 311–314, 1999.
3. E. Nöth, et al. VERBMOBIL: The use of prosody in the linguistic components of a speech understanding system. *IEEE Trans. on Speech and Audio Processing*, 8(5), 2000.
4. C. Schmandt. Employing voice back channels to facilitate audio document retrieval. *Proc. of COIS '88*, pp. 213–218, 1988.
5. T. Igarashi and J. F. Hughes. Voice as sound: Using non-verbal voice input for interactive control. *Proc. of UIST '01*, pp. 155–156, 2001.
6. E. Shriberg. To 'errrr' is human: ecology and acoustics of speech disfluencies. *J. International Phonetic Association*, 31(1):153–169, 2001.
7. T. Masui. An efficient text input method for pen-based computers. *Proceedings of CHI'98*, pp. 328–335, 1998.
8. K. Lyons, et al. Augmenting conversations using dual-purpose speech. *Proc. of UIST 2004*, pp. 237–246, 2004.