Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP-2004),
pp.1533-1536, October 2004.

Speech Spotter: On-demand Speech Recognition in Human-Human Conversation on the Telephone or in Face-to-Face Situations

Masataka Goto[†], *Koji Kitayama*[††], *Katunobu Itou*[‡], *and Tetsunori Kobayashi*[††]

[†] National Institute of Advanced Industrial Science and Technology (AIST). Ibaraki 305-8568, Japan
[††] Waseda University. Tokyo 169-8555, Japan [‡] Nagoya University. Aichi 464-8603, Japan
m.goto@aist.go.jp

Abstract

This paper describes a novel speech-interface function, called *"speech spotter"*, which enables a user to enter voice commands into a speech recognizer in the midst of natural human-human conversation. In the past, it has been difficult to use automatic speech recognition in human-human conversation since it was not easy to judge, from only microphone input, whether a user was speaking to another person or a speech recognizer. We solve this problem by using two kinds of nonverbal speech information: a filled pause (a vowel-lengthening hesitation like "er...") and voice pitch. Only when a user utters a voice command with a high pitch just after a filled pause is the voice command accepted by the speech recognizer. By using this speech-spotter function, we have built two application systems: an on-demand information system for assisting human-human conversation and a music-playback system for enriching telephone conversation. The results from using these systems have shown that the speech-spotter function is robust and convenient enough to be used in face-to-face or cellular-phone conversations.

1. Introduction

A computer system capable of speech recognition and information retrieval as if it was a particularly knowledgeable person would be a convenient aid to human-human conversation. When you, in conversation with another person, would like to obtain current information such as today's date and time, the weather forecast for today or tomorrow, or the results of sports events, you ordinarily would have to interrupt the conversation and ask yet another person or type a request into a computer to find the information you want. To eliminate the need for doing this, we aim at building a computer system which can monitor human-human conversations without disturbing them and provide information only when asked for it. This on-demand information assistance in human-human conversation will be useful and convenient because it will not be necessary to interrupt the original conversation to use some sort of input device (except for a microphone).

It has been difficult, however, to achieve a practical means of providing such assistance by using only microphone input. Previous approaches using only speech information detect keywords in speech signals by means of word-spotting technology [1, 2, 3]. These techniques, though, are poor at judging, without the context being restricted in advance, whether the detected keywords are intended to be *command utterances* for a computer system or *conversational utterances* for a conversational partner. To enable reliable judgment, previous speech-interface systems have had to use other input devices such as a button or a camera [4, 5]. In other words, no previous approach allowed a system to identify a *command utterance* in conversation without the context being restricted or some other device being used.

In this paper we describe a spotting function for speech input, called *speech spotter*, which enables a user to request speech-based assistance only when necessary while talking to another person. 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 such as "er..." or "uh..." (the lengthening of a vowel during hesitation). In other words, a computer system accepts this specially-designed unnatural utterance only and ignores other normal utterances in human-human conversation. For example, when a user asks the system by saying "Er... *(a filled pause)*, what's the date today? *(an utterance with a high pitch)*",[1] the system answers the question.

In the following sections, we explain the basic concept of speech spotter and describe the implementation of the speech-spotter function. We then introduce two practical applications and show that experimental results from twelve subjects have demonstrated the effectiveness of speech spotter.

2. Speech spotter

Speech spotter is a speech-interface function that enables a user to enter a word or sentence into a system through a speech recognizer whenever the user likes — even while talking to another person. We define an utterance spoken with the following two steps as a *speech-spotter utterance*:

1. Utter a filled pause by intentionally lengthening any vowel.
2. Utter a word or sentence by intentionally raising its pitch.

The system accepts this speech-spotter utterance and ignores other utterances and background noise. Because the speech-spotter utterance is unnatural and does not appear in natural conversation (at least not in Japanese),[2] the system does not interrupt normal human-human conversation even as it constantly monitors the conversation. For example, two people can use an information assistance system and a jukebox system with the speech-spotter function as illustrated in Figures 1 and 2.

The most important point is that a user can *intentionally* control whether each utterance is accepted (processed) by the speech recognizer. Since most current speech recognizers cannot judge whether a user is talking to another person or the system, all speech signals from the microphone are treated as utterances intended for the system. Typical spoken dialogue systems and computer telephony systems with speech recognizers therefore always assume that a user in front of the microphone is talking to the system and cannot be used to provide computer assistance while the user is talking to another person in a face-to-face situation or on the telephone. Although there were previous spotting approaches which required that an utterance intended for the system be preceded by a keyword — such as a general term like "Computer" or a system name like "Casper" or "Maxwell" — this restricted the usual behavior of a user: the user was forced to avoid use of the keyword in human-human conversation in front of the microphone. The speech-spotter function solves these problems by making use of intentional

[1] In this paper, underlining indicates that the pitch of the underlined words is intentionally raised by a user.

[2] Even if you may say "Uh..., what's the weather forecast?" in conversation, for example, you do not usually say "Uh..., what's the weather forecast?" with a high pitch.

Speaker A: Er..., what's the date today?

Speaker B: Let me see... I don't remember. Shall we ask the system? Er..., what's the date today? *(with a high pitch)*

The system then gives today's date through
a speech synthesizer or shows it on a screen.

Figure 1: *An example of using an information assistance system with the speech-spotter function.*

Speaker A: Let's change the background music.

Speaker B: How about Michael Jackson?

Speaker A: Uhm..., okay. I like his hit song from 1991.

Speaker B: Yeah, the title is "Black or White."

Speaker A: Uhm..., Black or White. *(with a high pitch)*

The system then plays the song "Black or White."

Figure 2: *An example of using a jukebox system with the speech-spotter function.*

nonverbal speech information, the filled pause and the voice pitch.

The speech-spotter function provides three benefits:

1. In human-human conversation, a speech recognizer can immediately be used whenever needed, as illustrated in Figures 1 and 2.
2. A hands-free interface system that does not require any input device other than a microphone is achieved. A user is free regarding body movement and can use the system even in telephone conversation.
3. A user can feel free to use any words in conversation with another person. The user does not have to carefully avoid saying anything that the system will accept as input.

3. Method of detecting speech-spotter utterances

Figure 3 shows a block diagram of the method for implementing the speech-spotter function: the four main processes are the *filled-pause detector*, *endpoint detector*, *speech recognizer*, and *utterance classifier*. The *speech recognizer* is implemented by modifying the CSRC (continuous speech recognition consortium) Japanese dictation toolkit [6] (*julian 3.3 beta* speech recognition engine [7] using a finite-state grammar).

In our method, speech-spotter utterances can be detected through the following four steps:

1. Each filled pause is detected by the *filled-pause detector* without using language information.
2. Upon being triggered by the detected filled pause, the *endpoint detector* determines the beginning of an utterance.
3. While the content of the utterance is being recognized by the *speech recognizer*, the end of the utterance is automatically determined by the *endpoint detector* on the basis of the intermediate recognition result.
4. The average pitch of the utterance whose beginning and end points were determined above is judged to be high or normal (low) by the *utterance classifier*. The high-pitch utterance is considered a speech-spotter utterance, while the normal-pitch utterance is simply ignored.

3.1. Detecting a filled pause *(filled-pause detector)*

To detect filled pauses in real time, we use a robust filled-pause detection method [8]. This is a bottom-up method that can detect a lengthened vowel in any word through a sophisticated signal-processing technique. It determines the beginning and end of each filled pause by finding two acoustical features of filled pauses — small fundamental frequency transitions and small spectral envelope deformations.

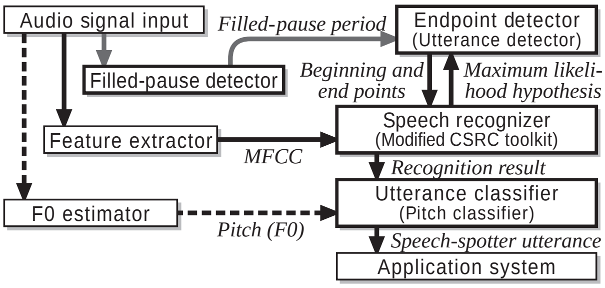

Figure 3: *Method for implementing the speech-spotter function.*

3.2. Determining the beginning of an utterance *(endpoint detector)*

Whenever a filled pause is detected, the beginning of the subsequent utterance is determined as being 130 ms before the end of the filled pause. Because this beginning point is thus determined as being in the middle of the filled pause, the *speech recognizer* can always start decoding at the stable vowel of the filled pause. This behavior is good for making the speech recognition stable. Every lengthened vowel (and subsequent consonant if necessary) should be inserted at the beginning of the grammar.

3.3. Determining the end of an utterance *(endpoint detector and speech recognizer)*

Although the beginning of an utterance is explicitly specified by the filled pause, the end of it is automatically determined by using an intermediate speech-recognition result, which is the maximum likelihood (ML) hypothesis in the HMM-based *speech recognizer*. After the *speech recognizer* starts decoding the current utterance, the *endpoint detector* checks the ML hypothesis for every frame. In the frame-synchronous Viterbi beam search, if the ML hypothesis stays at a unique node that is not shared by other words in a tree dictionary or a silence node that corresponds to the silence at the end of a sentence, its frame is considered the end of the utterance and the *speech recognizer* stops decoding [9]. In other words, it stops decoding when there is no possibility of other words or when it reaches the silence pause.

3.4. Judging the voice pitch *(utterance classifier)*

On the basis of a speaker-independent pitch-classification method using a threshold relative to the *base fundamental frequency (base F0)* [10], the *utterance classifier* filters out normal-pitch utterances to obtain high-pitch speech-spotter utterances. The base F0 is a unique pitch reference that corresponds to the pitch of the speaker's natural voice. After estimating the base F0 for each speaker, we can deal with the pitch value relative to the base F0, which compensates for a wide variety of voice pitch ranges. If the *relative pitch value* of an utterance, which is calculated by subtracting the base F0 from the pitch averaged over the utterance, is higher than a threshold, the utterance is judged to be a *speech-spotter utterance*; otherwise, it is discarded. The threshold is determined in advance to maximize the classification performance for a learning data set.

To estimate the voice pitch (the fundamental frequency) in real time, we use a sophisticated instantaneous-frequency-based analysis [8]: we find the most predominant harmonic structure in extracted frequency components by using a comb-filter-like analysis. As demonstrated by Goto *et al.* [10], the base F0 can be estimated by averaging the voice pitch during a filled pause. Since the filled pause is a natural hesitation that indicates a speaker is having trouble preparing a subsequent utterance and the speaker cannot change articulator states during filled pauses [8], the pitch during filled pauses is stable and is close to the pitch of the speaker's natural voice; i.e., the base F0.

4. Application systems

To demonstrate the usefulness of the speech-spotter function, we built two application systems: an on-demand information system for assisting human-human conversation, which achieves the information retrieval as illustrated in Figure 1, and a music-playback system for enriching telephone conversation, which achieves music playback as illustrated in Figure 2.[3]

4.1. On-demand information system for assisting human-human conversation

This system enables a user to retrieve information by saying the speech-spotter utterance of predefined query sentences when the user would like to obtain information while talking to another person. In our current implementation, the system supports the following simple information retrieval:

- Weather information
  In response to a query about the weather, such as "Uh..., the weather for tomorrow", the system shows the local weather forecast for today, tomorrow, the day after tomorrow, or for the next week. The forecast information is obtained on-the-fly from a WWW page.
- Date and time information
  In response to a query about the date and time, such as "Er..., what time is it?", the system shows today's date and time. This information is simply obtained from the operating system whose time is synchronized by the Network Time Protocol (NTP).
- News information
  In response to a query about the news, such as "Uhm..., sports news", the system shows the latest news headlines regarding world affairs, business, sports, and entertainment. This information is also obtained on-the-fly from a WWW page.

Other types of information retrieval can be easily supported by adding processes for interpreting the speech recognition results.

The retrieved results can be given to the user and the conversational partner through either a computer screen, a speech synthesizer, or both. There are forms of visual feedback for detecting a filled pause and a subsequent high-pitch utterance so that the user can understand how long a vowel should be lengthened and how high the voice pitch should be raised. The system can also be used in telephone conversation in combination with the music-playback system described next.

4.2. Music-playback system for enriching telephone conversation

This system enables a user to listen to background music while talking to another person by saying the speech-spotter utterance of the names of musicians and songs. Like the above on-demand information system, either a computer screen, speech synthesizer, or both can be used for the feedback. The system supports the following search methods:

- Specifying the song title
  When the song title is uttered, such as "Er..., Without You", the system plays back the appropriate song. It also shows the title on the screen or has the speech synthesizer read out the title.
- Specifying the artist's name
  When the artist's name is uttered, such as "Uhm..., Mariah Carey", the system shows a numbered list of song titles for that artist on the screen or has the speech synthesizer read out the list. After the user selects a song by saying the speech-spotter utterance of either the title or its number, the system plays back the song. It also highlights the selected title or reads out the title.

The system allows the user to say speech-spotter utterances at any time by overlapping and interrupting the speech synthesis or music playback. The user, for example, can stop the music playback by saying "Uh..., stop", or change the current song by saying another title.

This system is useful not only when a user would like to enjoy background music, but also when a user would like to talk about music in telephone conversations while listening to it; note that the system does not disturb such conversations. In particular, this is very effective for people who like to listen to music in everyday life because it makes it much easier for them to share background music and discuss it during playback on the telephone. The system can also be used to change background music in an actual room where people are talking, and can be used in addition to the above on-demand information system.

In our current implementation, all the names of musicians, songs, and the corresponding sound files are stored on an SQL (Structured Query Language) database server. Each musician and song name is also registered as a single word in the system vocabulary of the speech recognizer. We tested two different music databases, a database of 134 entries (names of 34 musicians and 100 of their songs) taken from the *"RWC Music Database: Popular Music"* (RWC-MDB-P-2001 Nos. 1−100) [11], and a database of 521 entries (names of 179 musicians and 342 of their songs) collected from Japanese hit charts during fiscal 2000.

5. Experimental results

We describe the results from evaluating the performance of the implemented speech-spotter function in Section 5.1 and then discuss the effectiveness of the two application systems in Section 5.2.

5.1. Evaluation of the speech-spotter performance

In this evaluation, we analyzed the rejection performance for normal utterances (Section 5.1.1) and detection performance for speech-spotter utterances (Section 5.1.2). While the speech-spotter function requires both the filled pause and the subsequent high-pitch utterance, a method using the latter, the high-pitch utterance only, could be considered (a method using the filled pause only is not practical because such pauses often appear in natural conversation). We therefore compared our method for detecting speech-spotter utterances with a method for detecting high-pitch utterances.[4]

5.1.1. Evaluation of rejection performance for normal utterances

We verified that normal conversational utterances would not be mistaken for command (speech-spotter) utterances. We tested both methods on 81 excerpts from the PASD spoken dialogue corpus [12], which consisted of normal utterances in conversation and naturally did not include speech-spotter utterances. Filled pauses appearing in each excerpt were used to estimate the base F0 for that excerpt. The total time for all the excerpts was about six hours.

Table 1 lists the number of mistakes where conversational utterances were detected as command utterances. As shown, the method of detecting only high-pitch utterances led to more mistakes than the speech-spotter method, so most mistakes can be eliminated by conditioning a user to add the preceding filled pause. The use of both the filled pause and the subsequent high-pitch utterance is thus necessary for high-precision detection.

5.1.2. Evaluation of detection performance for speech-spotter utterances

We also verified that speech-spotter utterances could be correctly detected. We tested both methods on an original corpus consisting of both normal and speech-spotter utterances by

[3] Video demonstrations of the two application systems are available at the following URL: http://staff.aist.go.jp/m.goto/ICSLP2004/

[4] To obtain each utterance for the method detecting high-pitch utterances, we used the typical endpoint-detection method with zero crossing rates and short time energy.

Table 1: *Comparison of the rejection performance: the number of mistakes where conversational utterances are detected as command utterances.*

	Voice pitch only	Speech spotter
# of mistakes	1338	60

Table 2: *Comparison of the detection performance: the recall and precision rates for detecting command (speech-spotter) utterances (218 words).*

	Voice pitch only	Speech spotter
Recall rate	0.78	0.78
Precision rate	0.35	0.77
Correctly detected	170 words	170 words
Substitution error	42 words	25 words
Deletion error	6 words	23 words
Insertion error	265 errors	3 errors

twelve Japanese subjects. We recorded speech-spotter utterances of 218 words (names of musicians and songs) and normal utterances of sentences naturally spoken with spontaneous filled pauses and coughs. The total time of the corpus was about 40 minutes.

Table 2 lists the recall and precision rates for detecting command (speech-spotter) utterances. The precision rate for the method of detecting only high-pitch utterances was low because there were many insertion errors. The use of the preceding filled pause in the speech-spotter method thus does not lower the recall rate and contributes to more robust detection.

5.2. Usability of speech-spotter-enabled application systems

In our experience with the two application systems, users having a spontaneous conversation were able to invoke the speech-recognition assistance by using the speech-spotter function without any training. Our on-demand speech assistance does not require that users switch a microphone on and off or use any other device, just that they say speech-spotter utterances, so the system was convenient and easy to use. The visual feedback of these systems led users to feel that the practical performance of detecting speech-spotter utterances was much higher than was reported in Section 5.1 because, for example, it was easy for the users to know how long a vowel should be lengthened during a filled pause.

5.2.1. Experience with the on-demand information system for assisting human-human conversation

Users chatting in front of the microphone were able to easily obtain information on the weather, date, and news through speech-spotter utterances. Whereas the use of a WWW browser for information searching in the midst of conversation was troublesome, this hands-free system was convenient and suitable for use in conversation. Some people found it interesting that the system responded only when an unnatural speech-spotter utterance was said, and enjoyed this behavior itself.

5.2.2. Experience with the music-playback system for enriching telephone conversation

Users were able to start playback of background music and change it while talking on cellular phones or normal phones. Although melody ringers (cellular phone ring-tones) are widely used, our users had no previous experience of listening to music in the midst of telephone conversation, and appreciated its novelty and usefulness. In our experience, the sound quality of music playback over cellular phones differed depending on the wireless carrier and phone model.

6. Conclusion

We have described a new speech-interface function *"speech spotter,"* which enables a system with a microphone only to judge whether an utterance in conversation is spoken to the system or another person. For the speech-spotter utterance, we use the unnaturalness of nonverbal speech information, an intentional filled pause and a subsequent high-pitch utterance, because this combination is not uttered in (Japanese) human-human conversation but is nevertheless easily uttered. We have also implemented two application systems, an on-demand information system to help users obtain information in human-human conversation, and a music-playback system to enable users to share music playback on the telephone as if they were talking in the same room with background music. As far as we know, this is the first system that people can use to obtain speech-based information assistance in the midst of a telephone conversation. Our experimental results have shown that the unnaturalness of nonverbal information can be used as a practical interface function.

The speech-spotter function is a general idea that enables a user engaged in conversation to enter voice commands into a system. We therefore plan to apply this idea to other voice-enabled applications. In addition, the concept of building a speech interface that uses nonverbal speech information intentionally controlled by a user originated from research on *"speech completion"* [13, 14], *"speech shift"* [10], and *"speech starter"* [9], which were followed by this research on *"speech spotter."* Our future work will also aim at further developing this concept.

7. References

[1] J. R. Rohlicek *et al.*, "Continuous hidden Markov modeling for speaker-independent word spotting," in *Proc. of ICASSP 89*, pp. 627–630, 1989.

[2] T. Kawahara *et al.*, "Speaking-style dependent lexicalized filler model for key-phrase detection and verification," in *Proc. of ICSLP 98*, pp. 3253–3256, 1998.

[3] R. E. Méliani and D. O'Shaughnessy, "Powerful syllabic fillers for general-task keyword-spotting and unlimited-vocabulary continuous-speech recognition," in *Proc. of ICSLP 98*, pp. 811–814, 1998.

[4] K. Nagao and A. Takeuchi, "Social interaction: Multimodal conversation with social agents," in *Proc. of AAAI-94*, vol. 1, pp. 22–28, 1994.

[5] Y. Matsusaka *et al.*, "Multi-person conversation via multimodal interface — a robot who communicate with multi-user," in *Proc. of Eurospeech '99*, pp. 1723–1726, 1999.

[6] T. Kawahara *et al.*, "Free software toolkit for Japanese large vocabulary continuous speech recognition," in *Proc. of ICSLP 2000*, pp. IV–476–479, 2000.

[7] A. Lee *et al.*, "Julius — an open source real-time large vocabulary recognition engine," in *Proc. of Eurospeech 2001*, pp. 1691–1694, 2001.

[8] M. Goto *et al.*, "A real-time filled pause detection system for spontaneous speech recognition," in *Proc. of Eurospeech '99*, pp. 227–230, 1999.

[9] K. Kitayama *et al.*, "Speech starter: Noise-robust endpoint detection by using filled pauses," in *Proc. of Eurospeech 2003*, pp. 1237–1240, 2003.

[10] M. Goto *et al.*, "Speech shift: Direct speech-input-mode switching through intentional control of voice pitch," in *Proc. of Eurospeech 2003*, pp. 1201–1204, 2003.

[11] M. Goto *et al.*, "RWC music database: Popular, classical, and jazz music databases," in *Proc. of ISMIR 2002*, pp. 287–288, 2002.

[12] S. Itahashi *et al.*, "Speech corpus by "Spoken Dialogue" project," in *Proc. of International Workshop on East-Asian Language Resources and Evaluation*, pp. 156–161, 1998.

[13] M. Goto *et al.*, "Speech completion: New speech interface with on-demand completion assistance," in *Proc. of HCI International 2001*, vol. 1, pp. 198–202, 2001.

[14] M. Goto *et al.*, "Speech completion: On-demand completion assistance using filled pauses for speech input interfaces," in *Proc. of ICSLP 2002*, pp. 1489–1492, 2002.