

Speech Spotter

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

Concept

New Direction of Speech Interface

- ❑ Exploit **nonverbal** speech information
 - Current speech-input interfaces have **not** fully exploited the potential of speech
 - Most speech recognizers utilize only **verbal** (phoneme) information

➔ Make use of nonverbal speech information **intentionally controlled** by a user



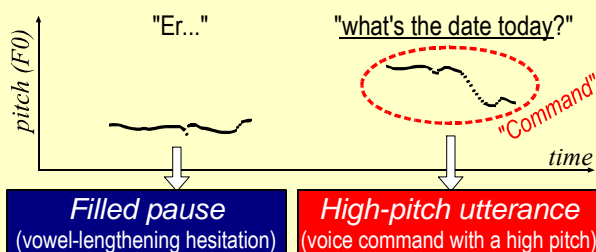
with nonverbal info.
(pitch and hesitation)



User-friendly
speech interface

Speech Spotter

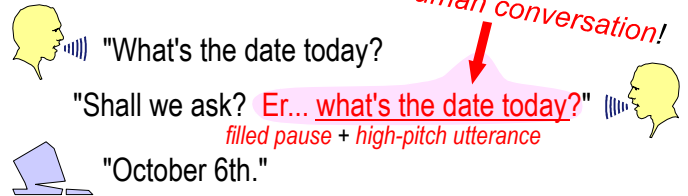
- ❑ What is speech spotter?
 - Enable a user to enter **voice commands** into a speech recognizer in the midst of **natural human-human conversation**
 - The system monitors human conversation and accepts only a **speech-spotter utterance** that does not appear in natural conversation
 - Use two kinds of **nonverbal speech information**



This is intentionally unnatural!

- Benefits
 1. Can immediately use a speech recognizer **whenever needed**
 2. Can use even in **telephone conversation**
Hands-free interface with only microphone
User is free regarding body movement
 3. Can **feel free to use any words** in conversation
Not have to carefully avoid saying anything that the system will accept

You can enter a **voice command** in the midst of **human-human conversation!**



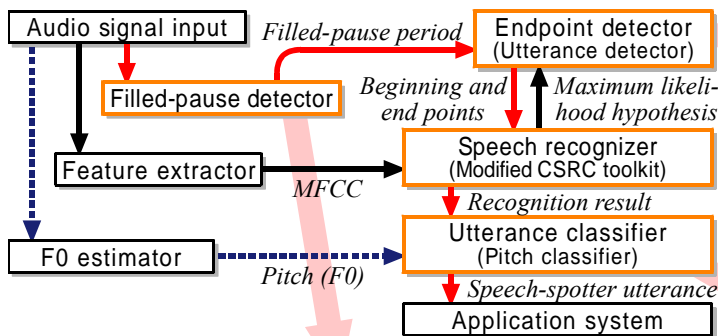
Previous Interfaces

- ❑ Difficult to use speech recognition in human-human conversation
 - Cannot judge whether a user is talking to **another person** or **the system**
 - Previous word-spotting approaches
Voice command is preceded by a **keyword** "Computer", "Casper", "Maxwell", etc.
User was forced to avoid use of the keyword
 - Use other input devices (button or camera)
The usual behavior of a user is restricted
Cannot be used in telephone conversation
- ➔ No practical means **using only microphone input**

Application Systems

- ❑ On-demand information system for assisting human-human conversation
 - **Obtain information** while talking to another person
System provide info. **only when asked for it**
 - **Weather** information (obtained from WWW)
Local weather forecast for today, tomorrow, etc.
 - **Date and time** information (obtained from NTP)
Today's date and time
 - **News** information (obtained from WWW)
Latest news headlines regarding world affairs, business, sports, and entertainment
- ❑ Music-playback system for enriching telephone conversation
 - **Listen to background music** while talking to another person **on the telephone**
 - Saying the **song title** (on SQL database server)
Play back the appropriate song (title on the screen)
 - Saying the **artist's name**
Show a numbered list of song titles for that artist
Play back the song selected by its title or number

Implementation

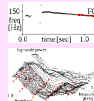


Endpoint Detector

- ❑ Detect the **beginning** of an utterance
 - Determine as being 130 ms before the FP end
- ❑ Detect the **end** of an utterance
 - Determine by checking the ML hypothesis
 - Stop when there is no possibility of other words
 - Stop when the hypothesis reaches the silence pause

Filled-Pause Detector

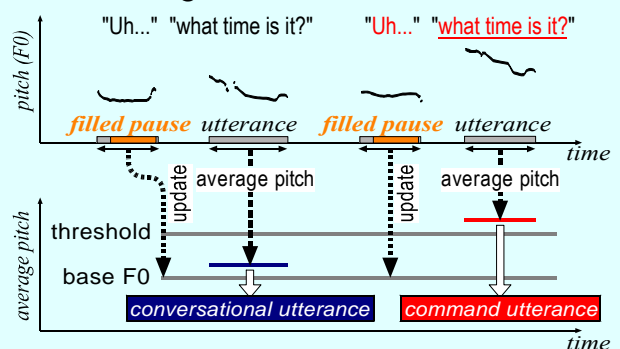
- ❑ Detect both ends of each filled pause
 - Real-time filled-pause (FP) detection method [Goto et al. 1999]
 - Independent of **vocabulary** and **language**
 - Detect a **lengthened vowel** in any word
 - Detect two acoustical features of FP
 - **Small pitch transition**
 - **Small spectral envelope deformation**



Utterance Classifier

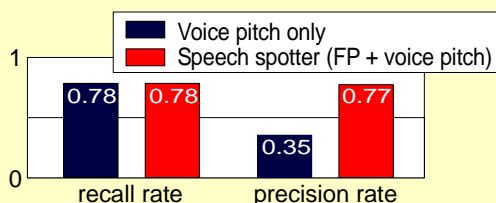
- ❑ Distinguish between **normal** and **high-pitch** utterances
 - Difficult to judge whether the pitch is raised
 - Pitch range differs among individuals
 - **Base fundamental frequency (base F0)**
 - Unique pitch reference for each speaker
 - Estimate by averaging the voice pitch during a **filled pause (FP)** (e.g., "uh...")
 - Use pitch value relative to the **base F0**

Thresholding method



Experimental Results

- ❑ Evaluation of detection performance
 - Compare the **speech-spotter function** with a **method w/o using the preceding FP**
 - Tested on 40-minute corpus consisting of normal and speech-spotter utterances
 - Uttered by 12 Japanese subjects
 - 218 speech-spotter utterances



➡ Use of preceding FP is important

- ❑ Usability of two application systems
 - Easy to use, can be used **without any training**
 - Convenient, **suitable for use in conversation**
 - Appreciate **novelty** and **usefulness**

First system that people can use to obtain **speech-based information assistance** in telephone conversation

Summary

- ❑ Propose a new speech interface function **"Speech Spotter"**
 - **Unnaturalness** of **nonverbal** speech info. can be used as a practical interface function
 - User can **intentionally control** whether each utterance is accepted (processed) by the speech recognizer
 - **General idea** can be used in other applications

Video Clips

Video clips of our speech-interface projects:
<http://staff.aist.go.jp/m.goto/SpeechCompletion/>
<http://staff.aist.go.jp/m.goto/ICSLP2004/>

On-demand information system for assisting human-human conversation

A: Hey, I've suddenly forgotten... What is the date today?
B: Yes, what is today's date? Well, shall we ask the On-Demand Conversation Assistance System? Er..., **what's the date today?**
* The system displays the current date and time:
"August 22, 2003, Friday, 23:51:10 JST"
A: Uh, it's already the 22nd!
B: Oh really? Well, that means our excursion is tomorrow. I hope it doesn't rain.
A: Shall we ask about the weather too? Er..., **what's tomorrow's weather?**
* The system checks tomorrow's weather report and displays the result:
"Clear"
B: Uh, no rain. Great!
A: That's good!



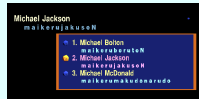
Music-playback system for enriching telephone conversation

B calls A on the telephone.
A: Yes...
B: Hello?
A: Uh..., what's up?
B: Thanks for all your help last time.
A: No problem. How have you been since?
B: Whew! I've been super busy writing that paper... I'm beat.
(Several minutes later)
A: Uh..., that reminds me, the song called "Fly Away" that we heard at that place, wasn't that good?
B: Oh, what song was that?
A: Shall we try listening to it?
B: What? We can hear it now?
A: Sure. This is a phone with a music-playback system. We can listen to that song like this... Er..., "Fly Away!"
* The system plays the song of that name on both of their handsets.
B: Wow, amazing! You can listen to a song by just saying its name!
Um..., this is a good song.
A: That's right!
(The conversation continues about various songs.)

Future Directions

□ Interfaces using intentional nonverbal info.

1. "Speech Completion" [HCI Intl. 2001] [ICSLP 2002]
2. "Speech Shift" [Eurospeech 2003]
3. "Speech Starter" [Eurospeech 2003]
4. "Speech Spotter" [ICSLP 2004]
5. "Speech ???"



Further developing this concept...

2004/10/06 ICSLP 2004 poster