Speech Shift

Direct Speech-Input-Mode Switching through Intentional Control of Voice Pitch

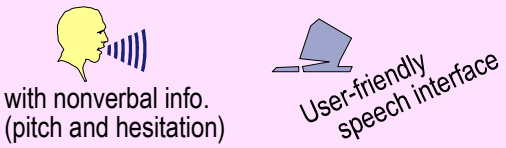
Concept

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

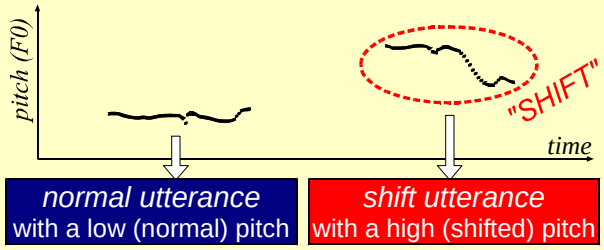
➡ Make use of nonverbal speech information intentionally controlled by a user



Speech Shift

❑ What is speech shift?

- Enable a user to switch speech-input modes by intentionally changing the pitch of an utterance
- Allocate two types of utterances to different modes



- Benefits
  1. Switching without other devices
     Can invoke functions in different modes w/o needing to use other devices
  2. Seamless switching between input modes
     Can invoke functions in different modes w/o switching between input modes explicitly w/o needing to be aware of the current input mode

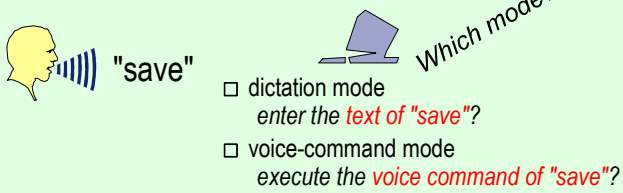
❑ General idea

- Useful for any voice-enabled applications
- Intentional pitch control brings additional information to speech-input interfaces

Previous Interfaces

❑ Word can be accepted in different modes

- Problem



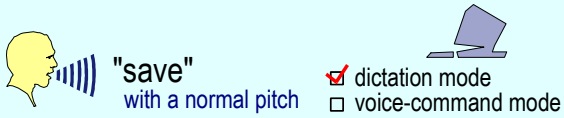
❑ Require explicit mode switching

- Use key phrases ("dictation" / "voice command")
  ➡ Hard to enter the phrases themselves
- Use other devices (mouse or keyboard)
  ➡ Awkward

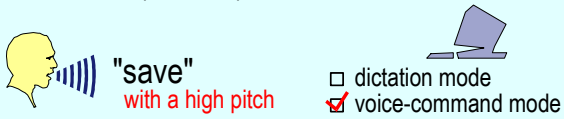
Voice-Enabled Word Processor

❑ Speech-shift function enables seamless speech-input-mode switching

- *normal utterance* : Dictation mode
  - Continuous speech dictation (text input)



- *shift utterance* : Voice-command mode
  - Edit-menu and format-menu commands
    delete, backspace, bold, left justify, right justify, center justify, new line (enter), undo, cut, paste, etc.
  - File-menu commands
    save, open (file open), print, close document, etc.

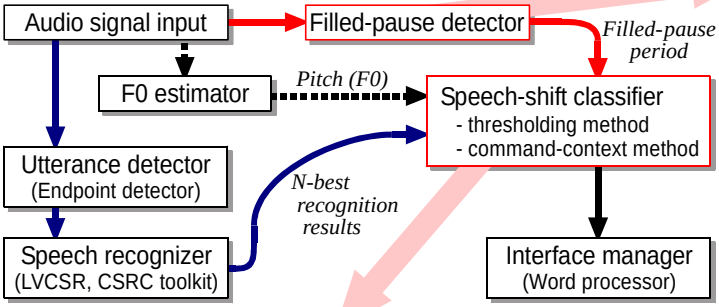


❑ Permit filled pauses

- Encourage a user to hesitate with filled pauses
  ➡ Utterances with filled pauses are not accepted as dictation

M. Goto [1], Y. Omoto [2], K. Itou [1], and T. Kobayashi [2]
[1] National Institute of Advanced Industrial Sci. and Tech. (AIST)
[2] Waseda University

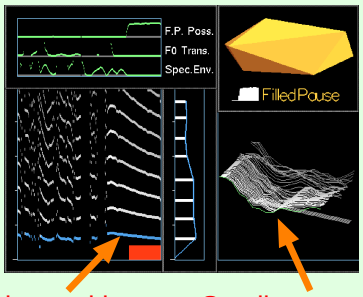
Implementation



Filled-Pause Detector

❑ Detect the beginning 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
- Bottom-up acoustical analysis
  - Two features of filled pause (FP)



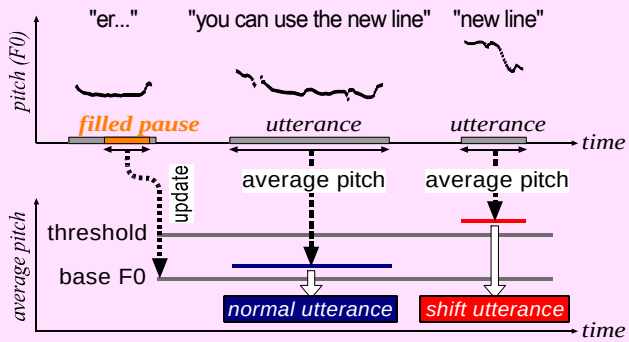
Small pitch transition — Small spectral envelope deformation

Speech-Shift Classifier

❑ Distinguish between normal and shift utterances

- Difficult to judge whether the pitch is shifted
  - 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., "er...")*
  - Gradually update the base F0 for every FP
- *Relative pitch value*
  - Pitch value relative to the base F0

❑ Thresholding method (for general purposes)



❑ Command-context method (for word processor)

- Incorporate prior knowledge about the linguistic context of voice commands

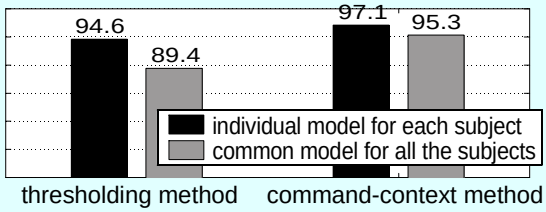
W: word sequence, ***C:*** command-flag sequence
X: spectrum sequence, ***A:*** pitch sequence

$$\{\hat{W}, \hat{C}\} = \underset{W, C}{\operatorname{argmax}}\ P(W, \boldsymbol{C} \mid X, \boldsymbol{A}) = \underset{W, C}{\operatorname{argmax}}\ \underbrace{P(\boldsymbol{A} \mid \boldsymbol{C})}_{\text{word-pitch model}}\ \underbrace{P(\boldsymbol{C} \mid W)}_{\text{command-flag model}}\ \underbrace{P(W \mid X)}_{\text{results of speech recognizer}}$$

Experimental Results

❑ Evaluation of classification performance

- Tested on 60 normal and 60 shift utterances
  - Uttered by 12 Japanese subjects



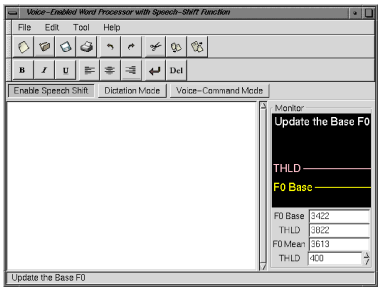
➡ Both methods are robust enough

❑ Usability evaluation of word processor

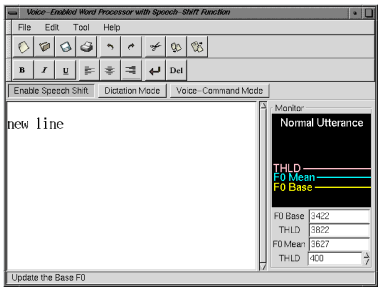
- Tested with 20 Japanese subjects
- Compare 4 input methods
  - (a) modes are switched by mouse operation
  - (b) switched by uttering predefined key phrases
  - (c) commands are uttered while pressing shift key
  - (d) commands are entered by speech-shift function
- Results
  - Required time: (a) and (d) took the shortest
  - Relative usage frequency: (d) was 79.8%
  - Questionnaire results: (d) was most preferred, easy to use, and labor-saving
  - 85% of the subjects wanted to use (d) in the future

Video clips of our speech-interface projects:
http://staff.aist.go.jp/m.goto/EUROSPEECH2003/
http://staff.aist.go.jp/m.goto/ICSLP2002/

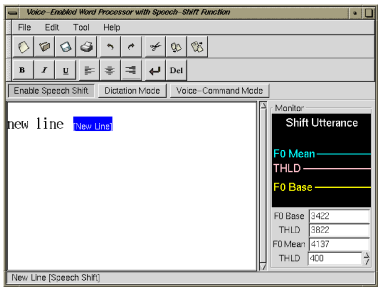
Snapshots



(1) Uttering "er...":
The *base F0* (the pitch of the speaker's natural voice) is updated whenever a filled pause is detected.



(2) Uttering "new line" with a normal pitch:
A *normal utterance* is regarded as regular dictation-mode text input.



(2) Uttering "new line" with a high pitch:
A *shift utterance* is regarded as voice-command-mode input.

Summary

❑ Propose a new speech interface function "Speech Shift"
- Make use of nonverbal speech info. (pitch)
- High-pitch voice has a good *"Shift"* function
  - Pitch of natural voice is estimated by using FPs
  - Naturally used in ventriloquism
- Effective means of entering voice commands
- Can be applied to various speech applications

Future Directions

❑ Interfaces using intentional nonverbal info.
- "Speech Completion" [HCI Intl. 2001] [ICSLP 2002]
- "Speech Shift" [Eurospeech 2003]
- "Speech Starter" [Eurospeech 2003]
- "Speech ???"

Further developing this concept...

2003/09/02 Eurospeech 2003 poster