

# **Speech-Recognition Interfaces for Music Information Retrieval: “Speech Completion” and “Speech Spotter”**

**Masataka Goto**

**Katunobu Ito**

**Koji Kitayama**

**Tetsunori Kobayashi**

AIST (National Inst. of Advanced Industrial Sci. & Tech.)

Nagoya University

Waseda University

Waseda University

*2004/10/13 ISMIR 2004*

# MIR + Speech Recognition = ?

## □ **Speech-Recognition Interface** is Well-Suited to MIR

- MIR-based jukebox system with speech-recognition interface

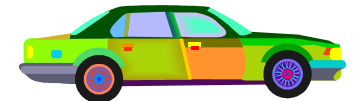
Retrieve a musical piece

just by saying the name of **musical piece** or **artist**

Change background music

at **home**

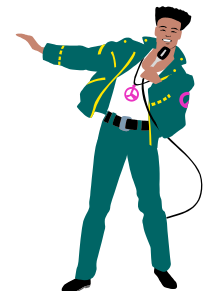
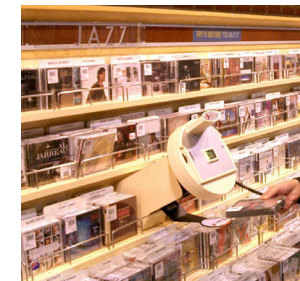
in a **car**



Find musical pieces

at music-listening stations in **music stores**

on **karaoke machines**





# Previous MIR Interface

---

## ❑ Text Query

- Enter **bibliographic information** by **keyboard** or **mouse**

## ❑ Example-based Query

- Find a **similar musical piece** by **(a) musical piece(s)**

## ❑ Melody-based Query

- Enter a **melodic contour** by **symbols**, **MIDI**, or **audio**
- **Query By Humming (QBH)** is promising

Require only a microphone

Can easily be used by a novice

## ❑ Speech Query

- **Not exploited how speech recognition can be used** for  
retrieving music information

# Our Contribution

## □ Two Original Speech-Recognition Interfaces

- Speech Completion
- Speech Spotter



## □ Two MIR-Based Jukebox Systems

- Music retrieval system with the Speech-Completion interface
- Music playback system with the Speech-Spotter interface

Find and play back a musical piece  
by saying its title or the artist's name



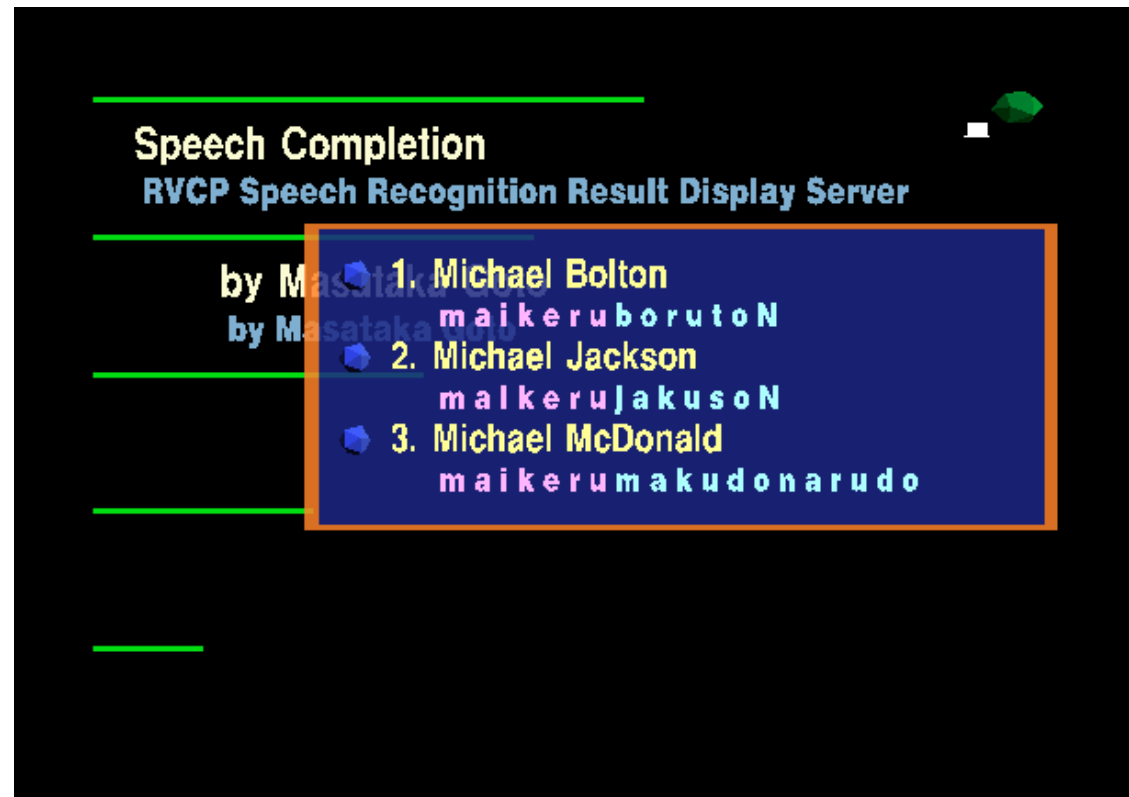
# Speech Completion

---

# Speech Completion

# Guess What's Happening?

❑ I'll Explain Later!

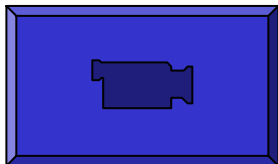


Speech Completion  
RVCP Speech Recognition Result Display Server

by M

- 1. Michael Bolton  
maikeruboruton
- 2. Michael Jackson  
maikerujakuson
- 3. Michael McDonald  
maikerumakudonarudo

The screenshot shows a dark interface with a list of suggestions. A blue box highlights the list. A small green icon is visible in the top right corner of the interface.



# Music Retrieval with Speech Completion

## ❑ The System Provides “*Completion*” Assistance

- You can retrieve a musical piece even if you **cannot remember** a part of the name

## ❑ Learning from Human-Human Conversation

You cannot remember a phrase “Michael Jackson” and **hesitate**

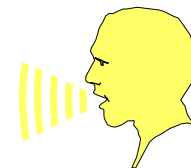
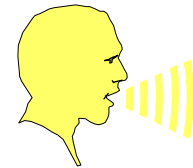
➡ A listener can **help you to recall it**

“Michael—” (Michael, uh...)

Filling in the rest of a fragment

||

**Completion**



“Michael Jackson?”

# Music Retrieval with Speech Completion

## ❑ The System Provides “Completion” Assistance

- You can retrieve a musical piece even if you cannot remember a part

## ❑ Learning from Human-Human

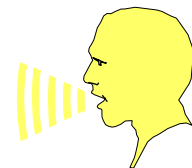
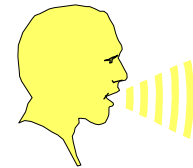
You cannot remember the rest of a fragment and hesitate you to recall it

“Michael—” (Michael, uh...)

the rest of a fragment

||

Completion



“Michael Jackson?”

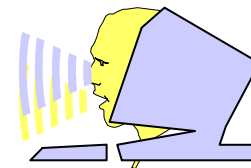
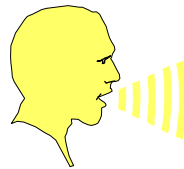
**Nobody proposed a “completion” function for speech interface!**

# Speech Completion

## □ What is **Speech Completion**?

- Help a user enter an uncertain piece/artist name by **completing the missing part** of a partially uttered fragment

“Michael—” (Michael, uh...)



“Michael Jackson?”

# Prevent Completion Becoming Annoying

## ❑ How to Invoke Speech Completion Function?

- Should be invoked **only when** a user **needs** an assistance



People **hesitates** when having trouble recalling

## ❑ Invoke by Using **Filled Pause (Lengthened Vowel)**

“Michael—” (Michael, uh...) ➔ “Michael Jackson?”

- **Filled pause** is a very natural trigger for human
- Can invoke the function **intentionally** and **effortlessly**

Frequently used in the same way in *Japanese* conversation

# Complete in Forward or Backward Direction

## ❑ Forward Speech Completion



- Do not remember the **last part** of a name

“Michael—” (Michael, uh...) → “Michael Jackson”

## ❑ Backward Speech Completion

- Do not remember the **first part** of a name



“Nantoka— (Something—) Jackson” → “Michael Jackson”  
“Janet Jackson”

Wildcard keyword

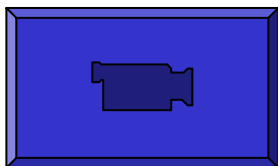
# System Output (Video Demonstration)

## □ Demonstration of Speech Completion

- Enter the *Japanese* names of **musicians** and **songs**

“Michael Jackson”  
↓  
“MAIKERU JAKUSON”  
(in Japanese)

“Michael—”  
↓  
“MAIKERU—”



Speech Completion  
RVCP Speech Recognition Result Display Server

by Masataka

1. Michael Bolton  
maikeruboruton
2. Michael Jackson  
maikerujakuson
3. Michael McDonald  
maikerumakudonarudo



# Advantages of Speech Completion

---

## □ Provide Three Benefits

1. A user can more easily recall poorly remembered names
2. Less labor is needed to input a long name

*“Supercalifragilisticexpialidocious”* (a song from “*Mary Poppins*”)

3. The user is not forced to utter the entire name  
carefully and precisely

## □ Can Be Used for Various MIR Situations

- Playing back a SMF or *karaoke* track
- Making a play list
- Editing and downloading music files



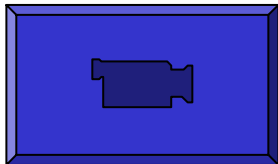
# Speech Spotter

---

# Speech Spotter

# Guess What's Happening?

- ❑ I'll Explain Later!



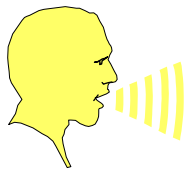
**High-pitch utterance**



Er..., "Fly Away"!

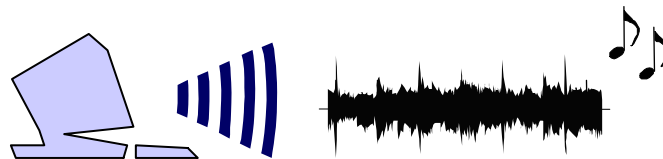
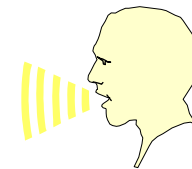
# Music Playback with Speech Spotter

- ❑ **On-Demand Music Playback Assistance**  
in **Human-Human Conversation**
  - You can listen to background music by saying its **title** or the **artist's** name **while talking to another person**



“Shall we listen to the song ‘Black or White’ ?”

“Yeah! Uhm... Black or White.”





# How To Achieve On-Demand Assistance

---

## ❑ Problem To Be Solved

- Monitor human-human conversations  
without disturbing them
- Provide music playback **only when asked** for it

## ❑ Difficult To Achieve By Using **Only Microphone**

### **✗** Word-spotting technology

Poor at judging whether the detected keywords are  
**Command utterances** for a computer system or  
**Conversational utterances** for a conversational partner

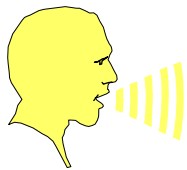
### **✗** Using button or camera

Require **input devices** other than a microphone  
**Cannot** be used in a **telephone** conversation

# Speech Spotter

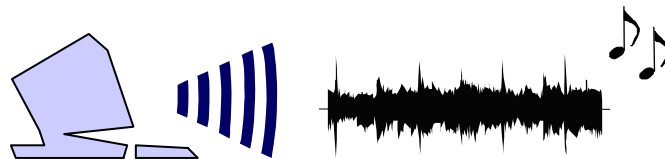
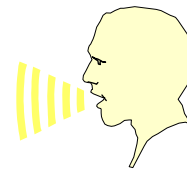
## ❑ What is **Speech Spotter**?

- Regard a user utterance as a **command utterance** only when it is **intentionally** uttered with a **high pitch** just after a **filled pause** (e.g., “er...”)



“Shall we listen to the song ‘Black or White’ ?”

“Yeah! Uhm... **Black or White.**”





# Speech Spotter

---

## ❑ What is **Speech Spotter**?

- Regard a user utterance as a **command utterance** only when it is **intentionally** uttered with a **high pitch** just after a **filled pause** (e.g., “er...”)

This combination is quite **unnatural**

The system accepts

this **specially-designed unnatural** utterance only

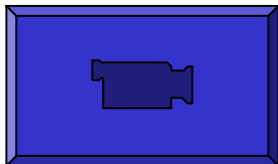
This does **not appear** in natural conversation

➡ The system can easily **spot** it!

# System Output (Video Demonstration)

## ❑ Demonstration of Speech Spotter

- Enter voice commands for **music-playback control**





# System Implementation

---

## ❑ Implementation of Two Interfaces

- Several difficulties should be overcome

## ❑ New Technology

- **Real-time filled-pause detection**

Detect a **lengthened vowel** in any word

- **Completion-candidate generation**

Generate **candidates** by tracing on a vocabulary tree

- **Endpoint detection**

Detect the **beginning** and **end** of an utterance

- **Utterance-pitch classification**

Distinguish between **normal** and **high-pitch** utterances



# Conclusion

---

## ❑ Summary

- MIR system with **Speech Completion**  
Listen to a piece even if part of its name cannot be recalled
- MIR system with **Speech Spotter**  
Share music playback on the telephone

## ❑ Practical Speech-Recognition Interfaces for MIR

- **Cannot** be achieved by simply applying  
the **current ASR** (automatic speech recognition) to MIR

## ❑ First Step Toward Building **Ultimate MIR Interface**

- Important to explore various speech-recognition interfaces  
for MIR

# Future Directions

## □ QBH + ASR

- Build a **unified system** using a microphone input
  - query-by-humming systems
  - speech-recognition interfaces
- Leverage the **potential affinity** between them



+



+

