Automatic Discrimination between Singing and Speaking Voices for a Flexible Music Retrieval System

Yasunori Ohishi [1] *Masataka Goto* [3]
Katunobu Itou [2] *Kazuya Takeda* [1]

[1]Nagoya University, Japan

[2] Hosei University, Japan

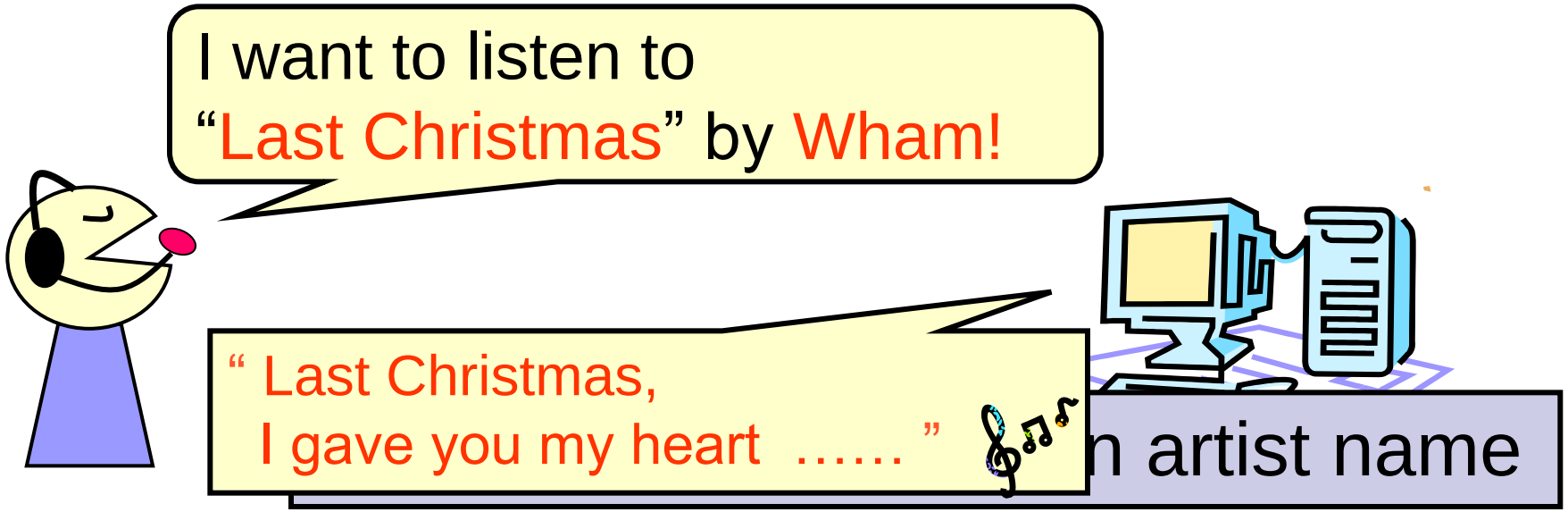
[3] National Institute of Advanced Industrial Science and Technology (AIST)

Existing music retrieval system

I want to listen to
“Last Christmas” by Wham!

Saying a song title or an artist name

Existing music retrieval system



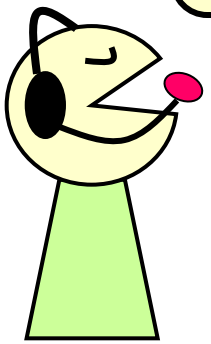
Existing music retrieval system

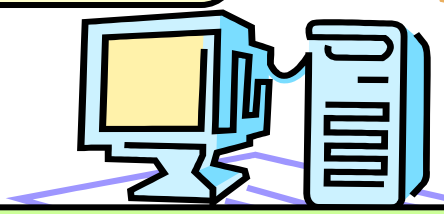
I want to listen to
"Last Christmas" by Wham!

" Last Christmas,
I gave you my heart ……"

n artist name

La ~~
La ~~





Singing or humming a song melody

Existing music retrieval system

I want to listen to
"Last Christmas" by Wham!

" Last Christmas,
I gave you my heart ……"

n artist name

La ~~
La ~~

" Last Christmas,
I gave you my heart ……"

song melody

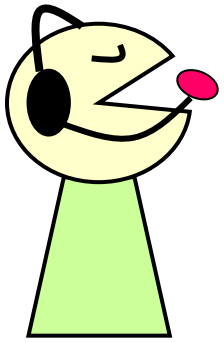
Proposed music retrieval system

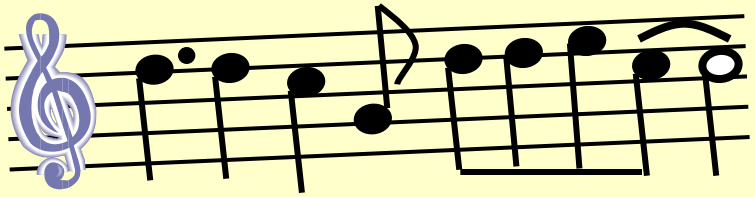
I want to listen to
"Last Christmas" by Wham!

" Last Christmas,
I gave you my heart ……"

Input voice
II
Speaking

" Last Christmas,
I gave you my heart ……"





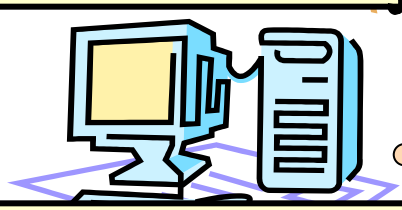
La ~~
La ~~

Proposed music retrieval system

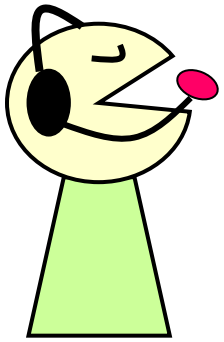
I want to listen to
"Last Christmas" by Wham!

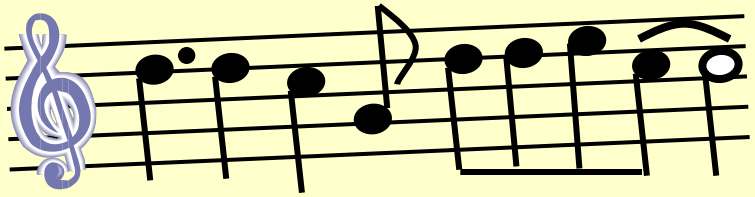
" Last Christmas,
I gave you my heart ……"

Input voice
II
Singing



" Last Christmas,
I gave you my heart ……"





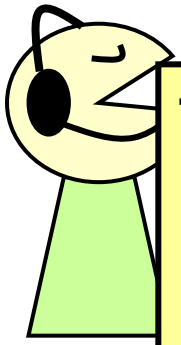
La ~~
La ~~

Proposed music retrieval system

I want to listen to
"Last Christmas" by Wham!

" Last Christmas,
I gave you my heart ……"

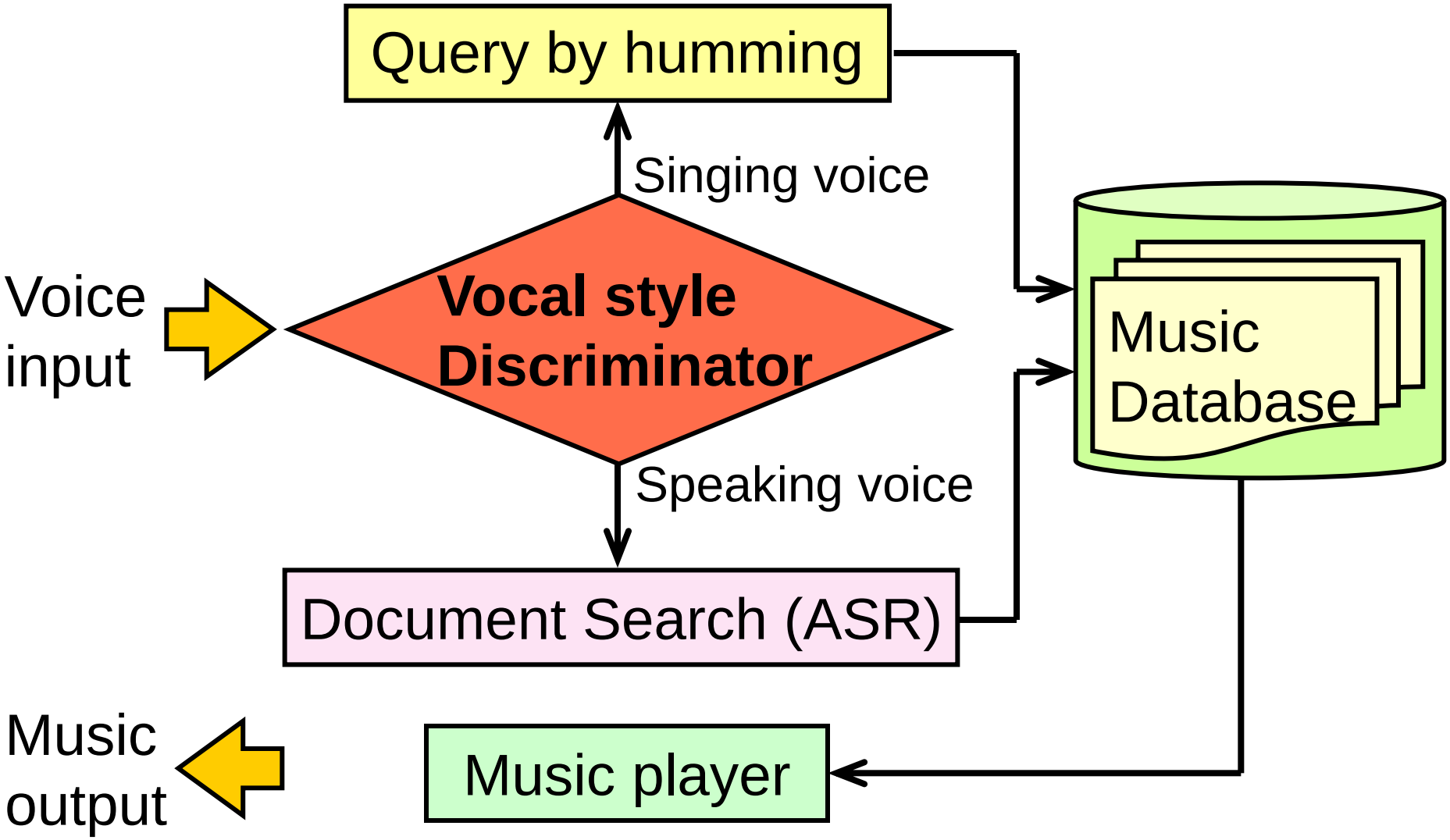
" Last Christmas,
I gave you my heart ……"



This system can accept
both **singing voices for the melody**
and **speaking voices for the title**

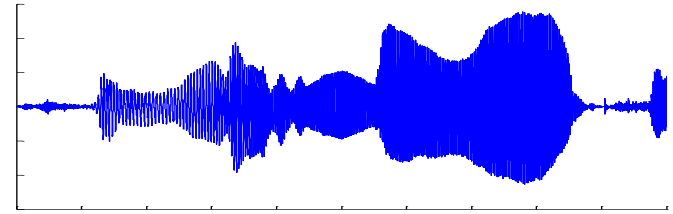
Let’s watch a demo video for this system

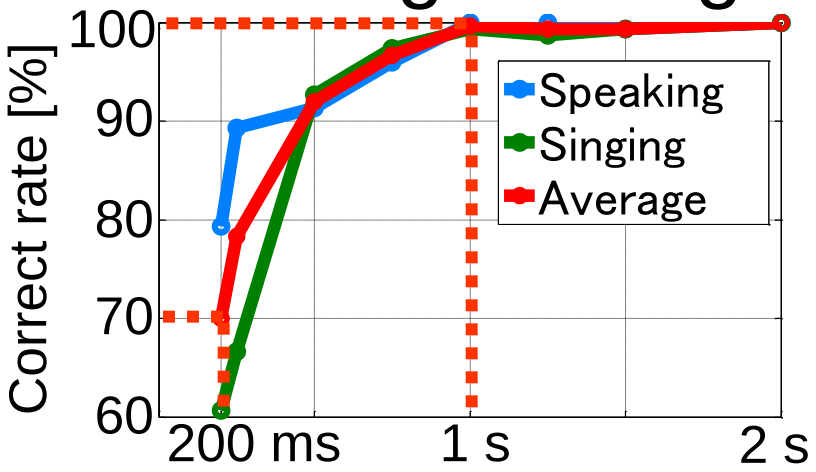
Flexible Music Retrieval System



Subjective experiments

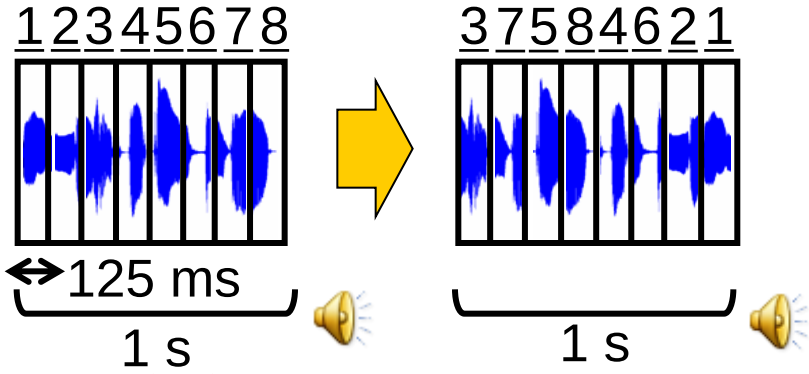
- Necessary to investigate voice signal length





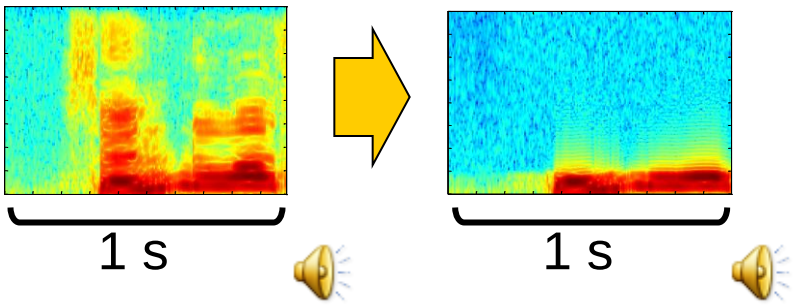
- Necessary to investigate acoustic cues

Random splicing technique



✕ Temporal structure

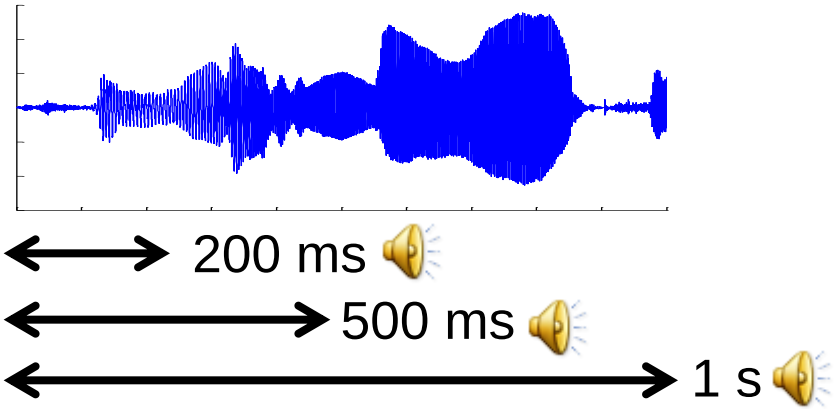
Low-pass filtering technique

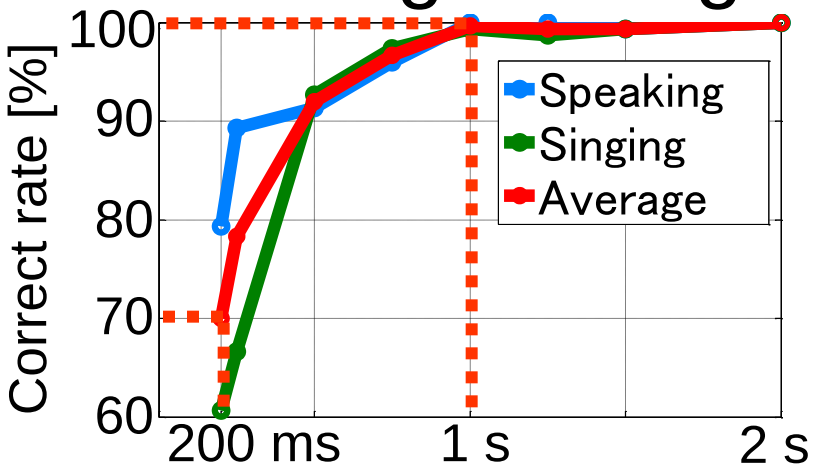


✕ Short-time spectral feature

Subjective experiments

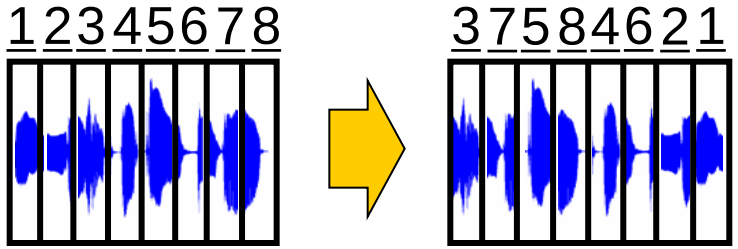
- Necessary to investigate voice signal length



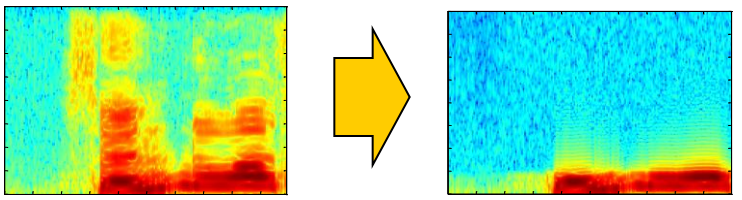


- Necessary to investigate acoustic cues

Random splicing technique



Low-pass filtering technique



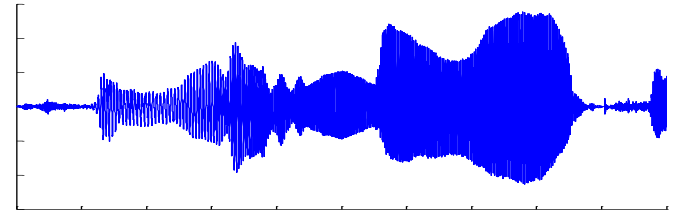
Speaking voices → Speaking voices ○

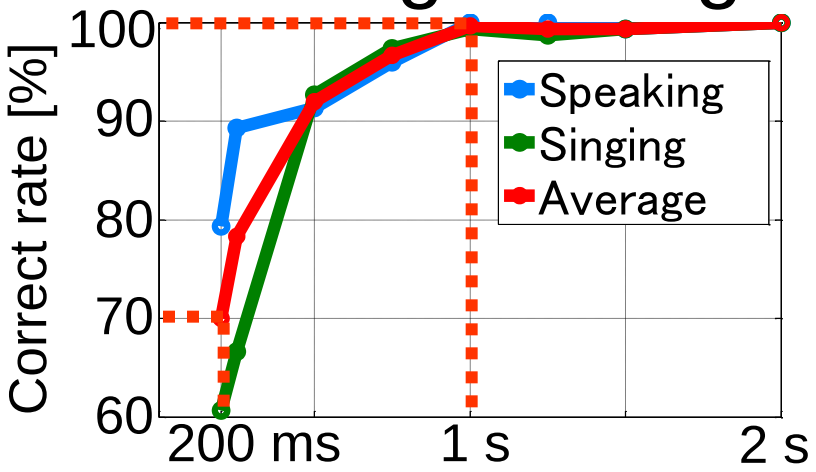
Singing voices → Speaking voices ×

tral feature

Subjective experiments

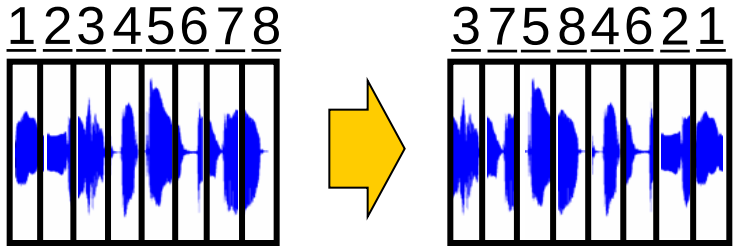
- Necessary to investigate voice signal length



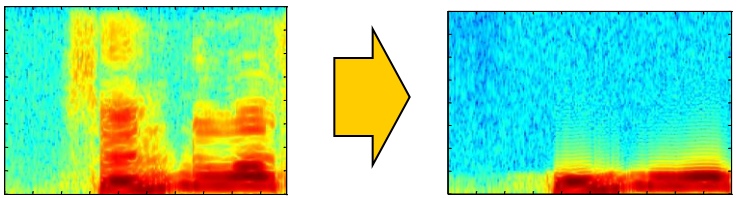


- Necessary to investigate acoustic cues

Random splicing technique



Low-pass filtering technique

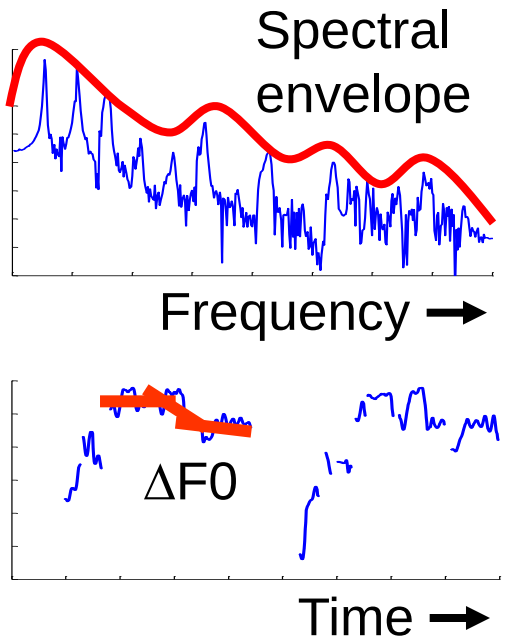


Prosodic and spectral cues complementarily contribute to perceptual judgments

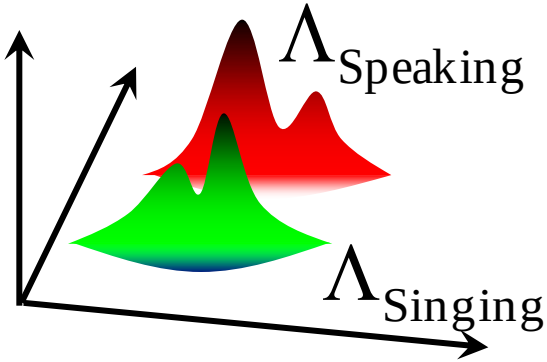
al feature

Discrimination measures

- Spectral feature measure
  - Mel-Frequency Cepstrum Coefficients (**MFCC**)
  - **ΔMFCC** (5-frame regression)
- Dynamics of prosody
  - **ΔF0** (5-frame regression)
    - F0 Extraction (*PreFEst*, Goto1999)
- Training the discriminative model
  - **MFCC+ΔMFCC+ΔF0**
  - Gaussian Mixture Model (GMM)



$$\hat{d} = \arg\max_{d=\text{Singing},\text{Speaking}} \frac{1}{T}\sum_{t=1}^{T} \log p(\mathbf{x}_t; \Lambda_d)$$



Experimental evaluation

- AIST humming database
  - Sing and hum chorus and verse A sections
    - At an arbitrary tempo, without musical accompaniment
  - Read the lyrics of chorus and verse A sections

75 Japanese subjects (25 Japanese songs)

	Humming	Singing	Speaking
Chorus	1875 samples	1875 samples	1875 samples
Verse A	1875 samples	1875 samples	1875 samples

25 native English speakers (10 western songs)

	Humming	Singing	Speaking
Chorus	250 samples	250 samples	250 samples
Verse A	250 samples	250 samples	250 samples

Experimental evaluation

- AIST humming database
  - Sing and hum chorus and verse A sections
    - At an arbitrary tempo, without musical accompaniment
  - Read the lyrics of chorus and verse A sections

75 Jap

Test data

GMM training data

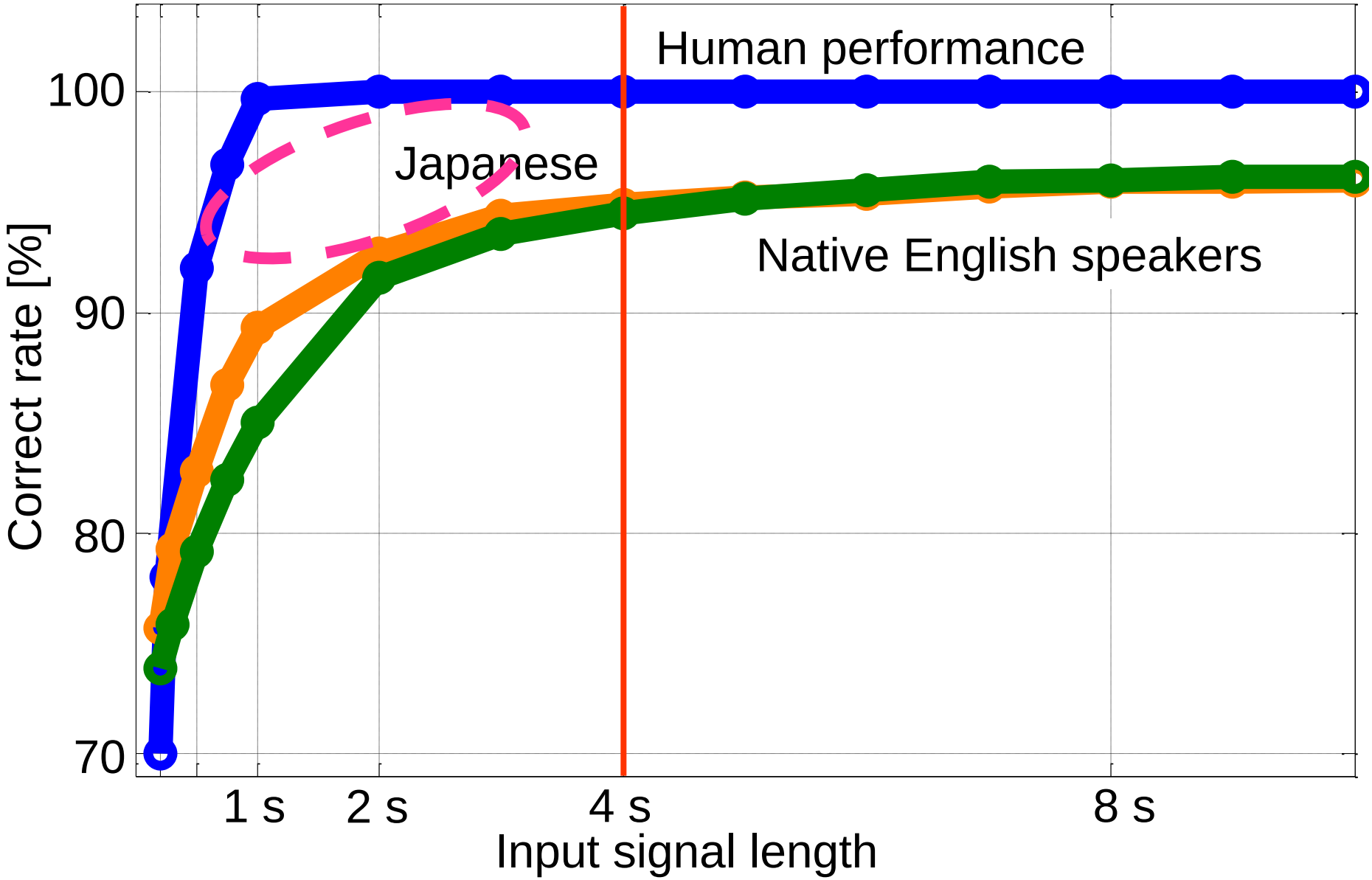
Chorus	1875 samples	1875 samples	1875 samples
Verse A	1875 samples	1875 samples	1875 samples

25 nativ ers (10 western songs)

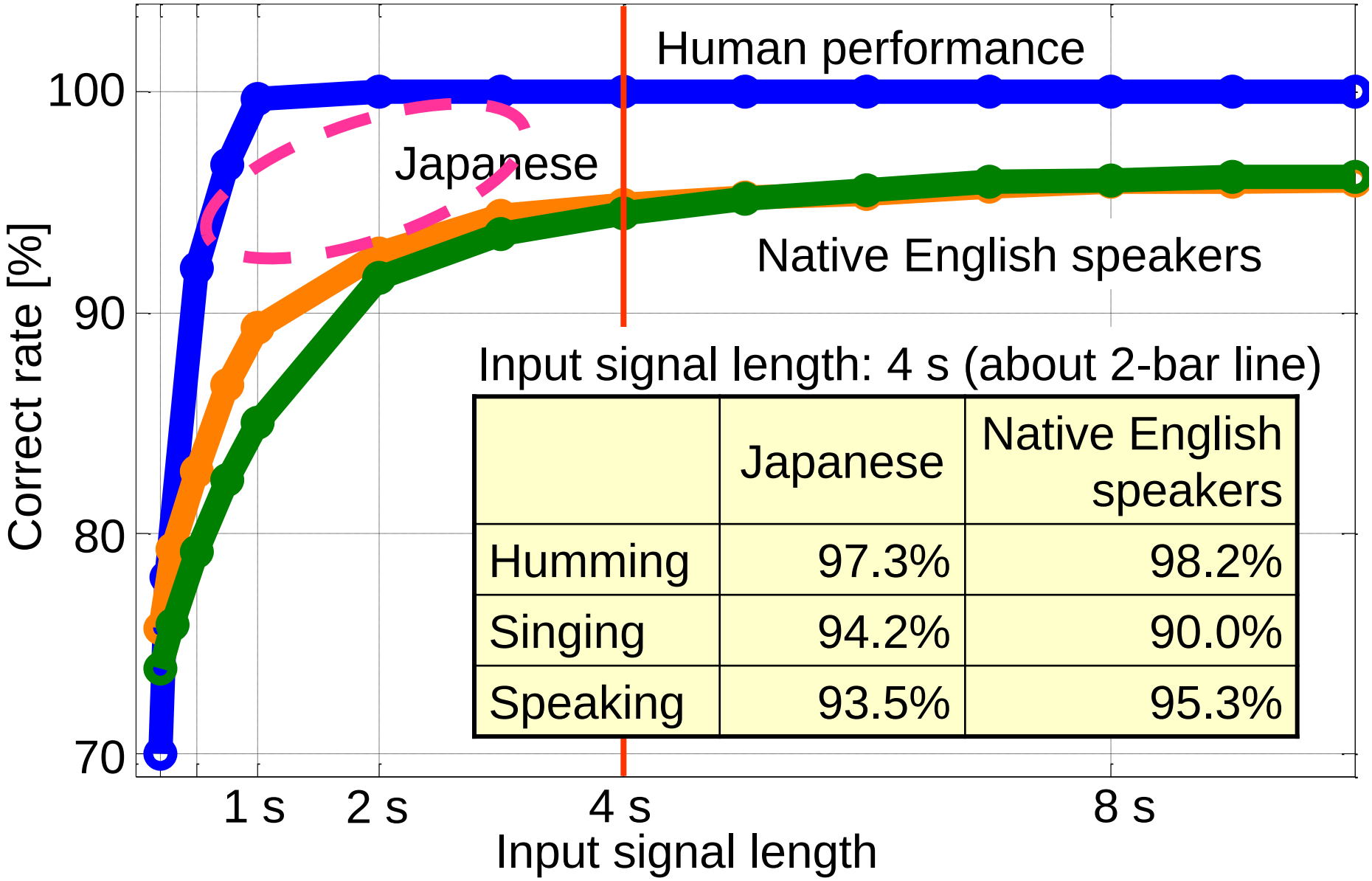
Test data

		Singing	Speaking
Chorus	250 samples	250 samples	250 samples
Verse A	250 samples	250 samples	250 samples

Discrimination results



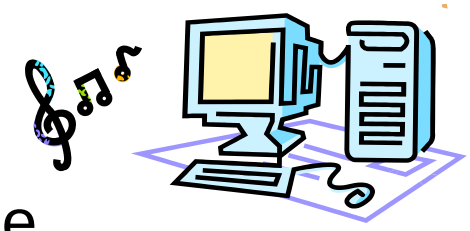
Discrimination results



Input signal length: 4 s (about 2-bar line)

	Japanese	Native English speakers
Humming	97.3%	98.2%
Singing	94.2%	90.0%
Speaking	93.5%	95.3%

Conclusion



- Flexible music retrieval system
  - Users can retrieve a song by singing the melody or saying the title
- Automatic discrimination between singing and speaking voices
  - Prosodic and spectral cues
  - Discrimination measures → MFCC, F0 contour
  - 94.7% correct rate obtained for 4-s signals
- New measures for automatic discrimination
- Evaluation experiments
  - Query-by-humming method
  - Automatic speech recognition

Thank you for your attention

Any question ?