On Human Capability and Acoustic Cues for Discriminating Singing and Speaking Voices

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

[1] Graduate School of Information Science, Nagoya University, Japan

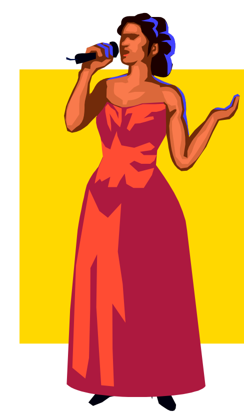
[2] Faculty of Computer and Information Sciences, Hosei University, Japan

[3] National Institute of Advanced Industrial Science and Technology

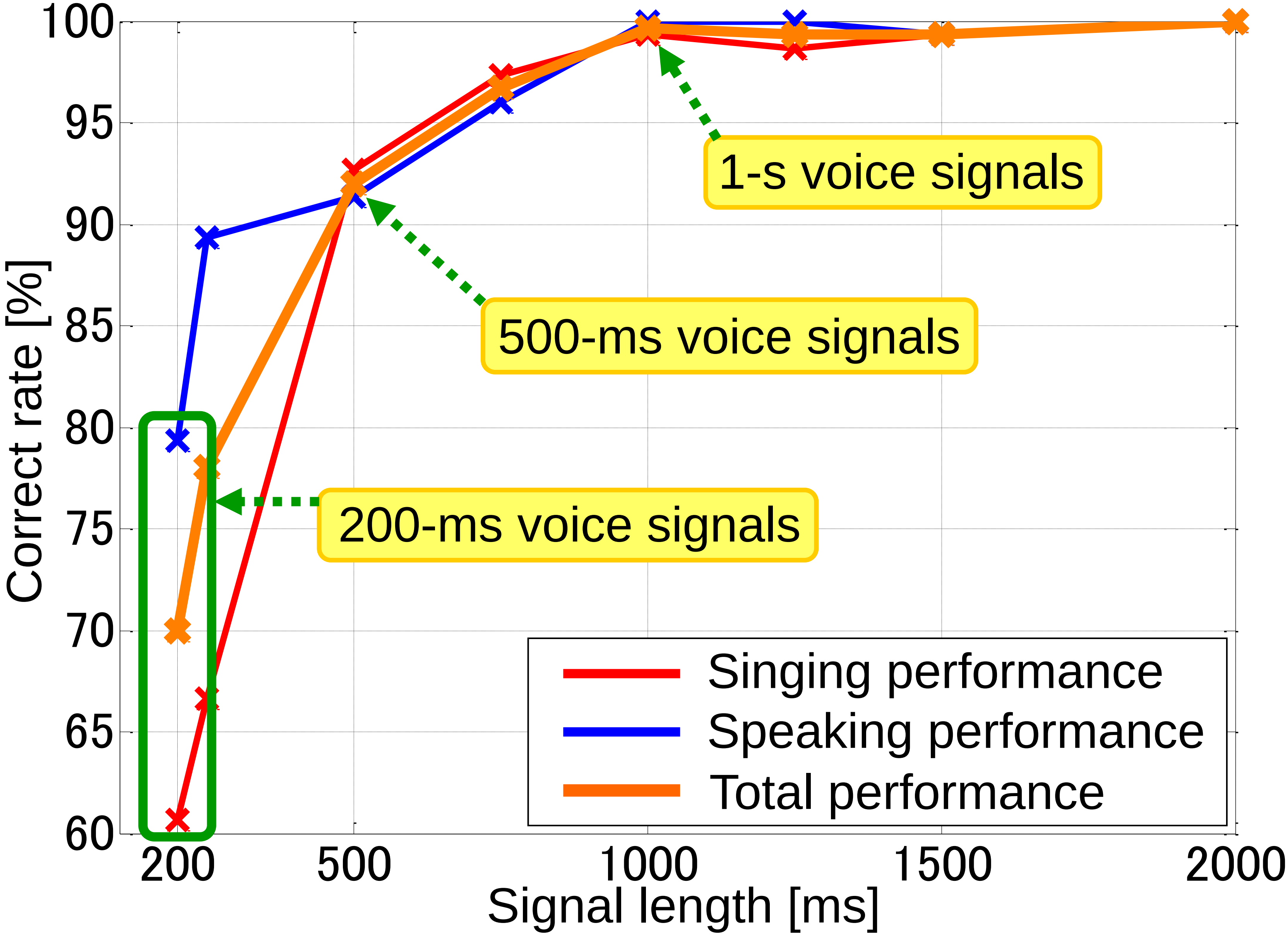
Let’s do the Quiz

- Can you discriminate between **Singing** and **Speaking** voices?
  (Japanese voices)

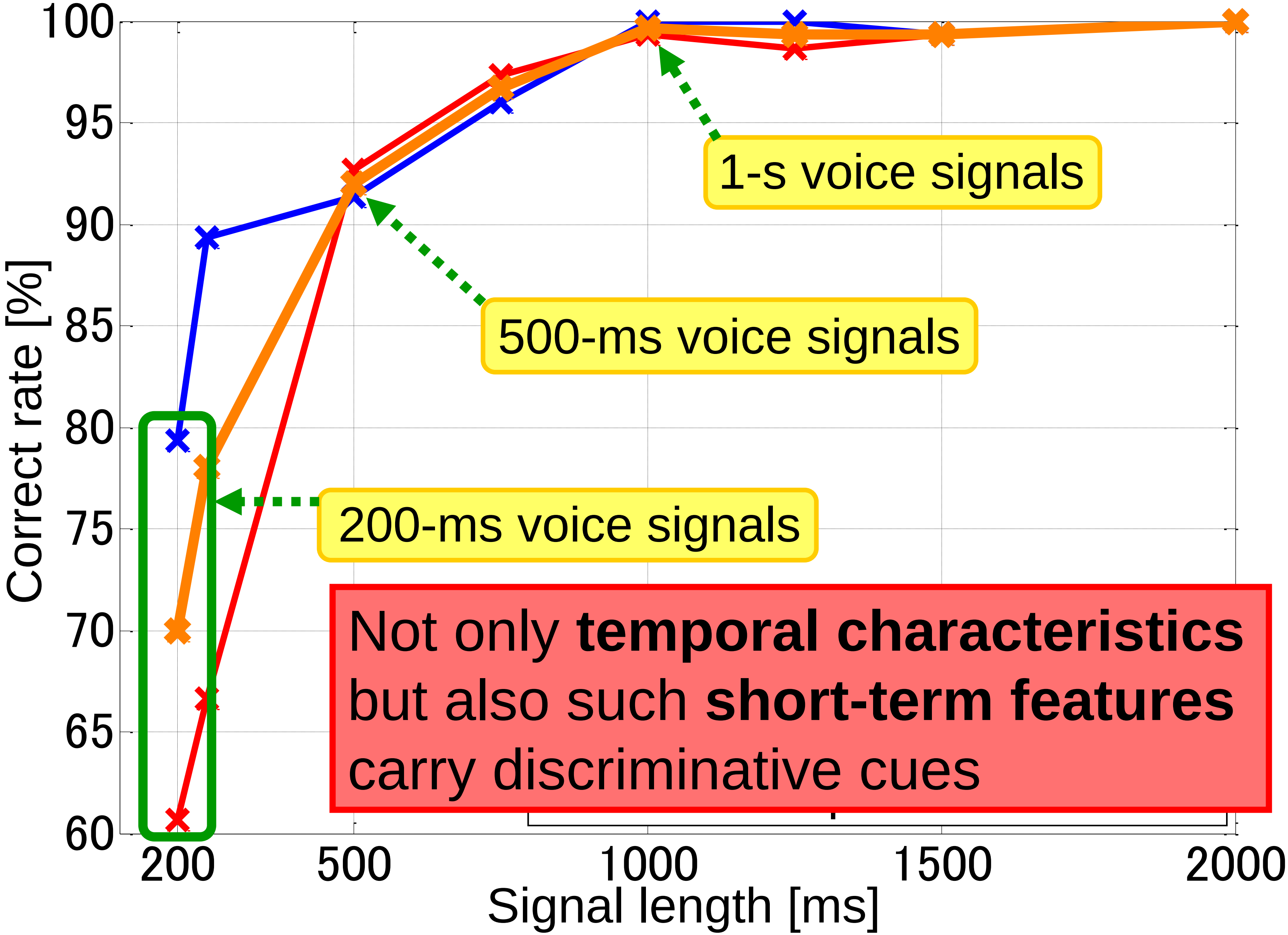
Q.1. Can you do it ? (2 s long)	
Q.2. Can you do it ? (500 ms long)	
Q.3. Can you do it ? (200 ms long)	



Investigation of signal length necessary for discrimination



Investigation of signal length necessary for discrimination



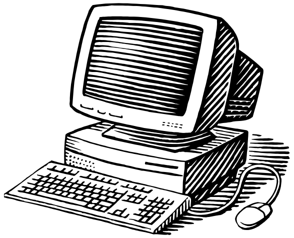
The goal of this study

Subjective experiments

Investigation of acoustic cues necessary for discrimination between singing and speaking voices

Based on knowledge obtained by subjective experiments

Automatic vocal style discriminator

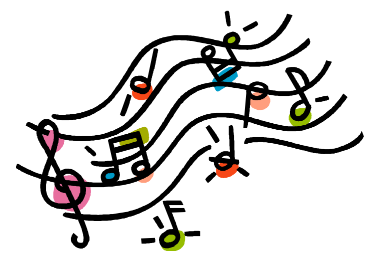


- Spectral feature measure
- F0 derivative measure

Introduction of the voice database

- AIST humming database
  - 75 Japanese subjects（37 males, 38 females）
  - **Sing a chorus and verse A sections**
    at an arbitrary tempo,
    without musical accompaniment
    (25 Japanese songs selected from
    “RWC Music Database: Popular Music”)
  - **Read the lyrics of chorus and verse A sections**

    Most of these subjects haven’t had the special musical training



The goal of this study

Subjective experiments

Investigation of acoustic cues necessary for discrimination between singing and speaking voices

Based on knowledge obtained by subjective experiments

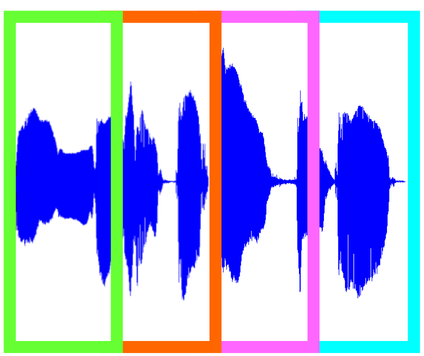
Automatic vocal style discriminator

- Short-term spectral feature measure
- F0 derivative measure

Investigation of acoustic cues necessary for discrimination

Temporal structure of signal is modified,
short-time spectral features are maintained

<u>Random splicing technique</u>



250 ms

1 s

Randomly concatenating pieces

Let's do the quiz

Q.1 (250 ms)

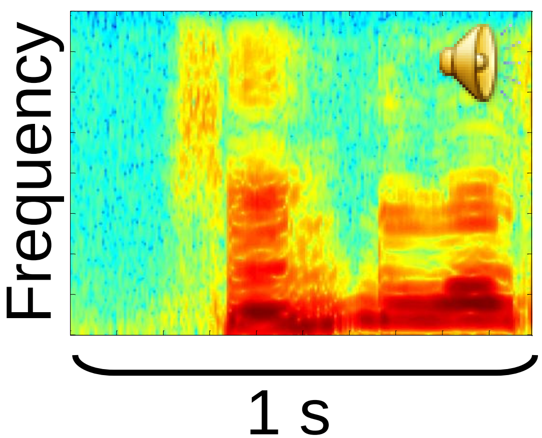
Q.2 (200 ms)

Q.3 (125 ms)

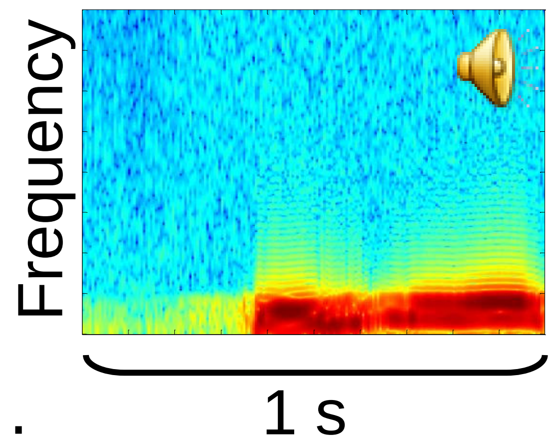
Investigation of acoustic cues necessary for discrimination

Temporal structure of signal is maintained,
short-time spectral features are modified

<u>Low-pass filtering technique</u>



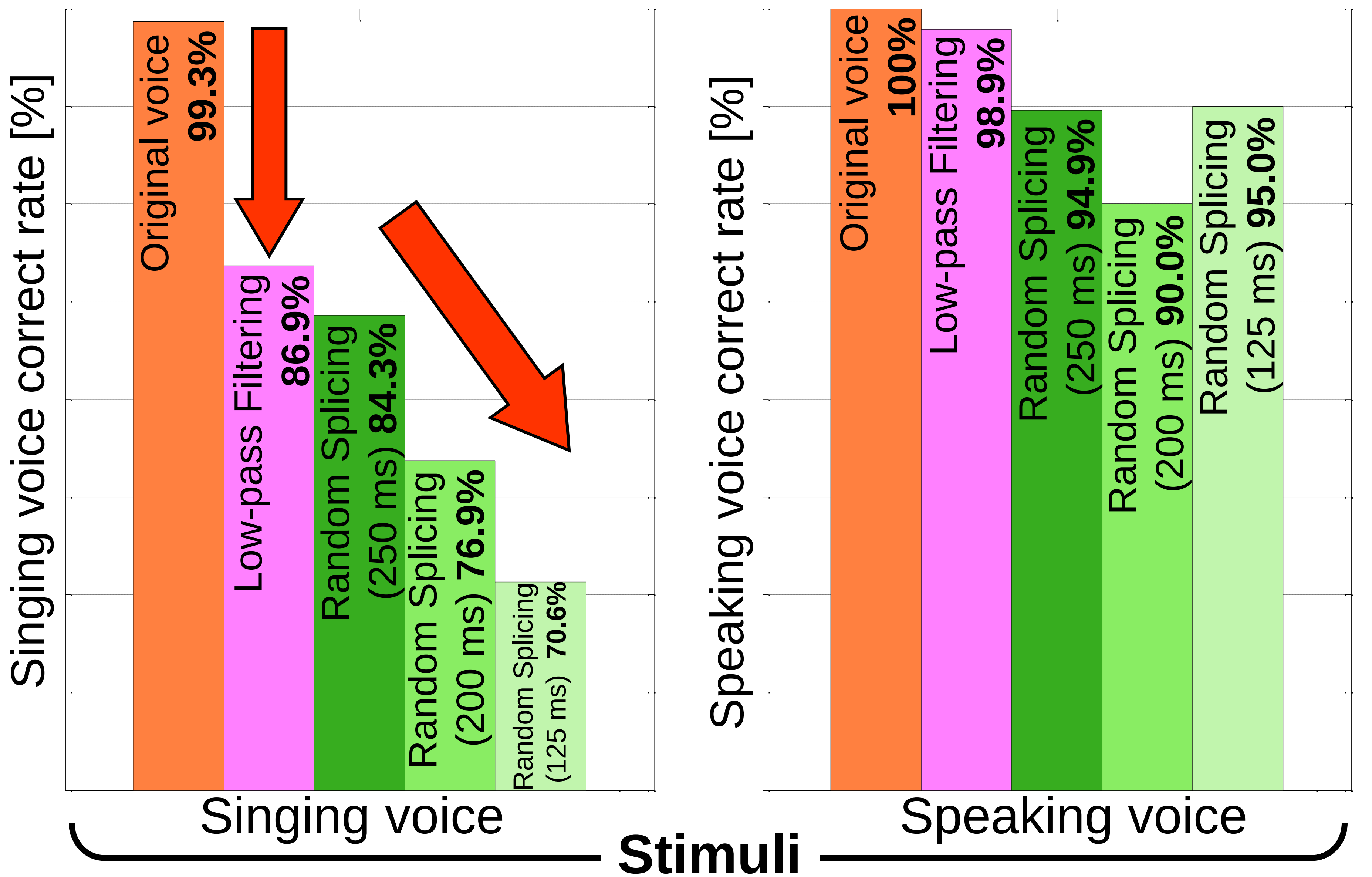
Eliminating frequency component higher than 800 Hz



Let’s do the quiz

Q.1 Q.2 Q.3

Investigation of acoustic cues necessary for discrimination



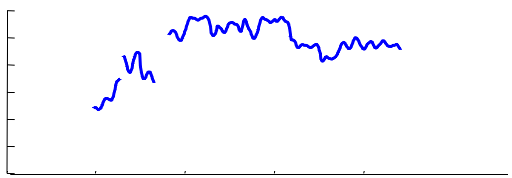
Discussion

- Correct rate of singing voices declined

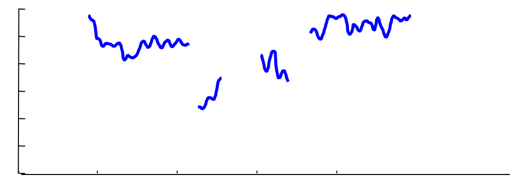
Random splicing technique

- Temporal structure of the original voices (rhythm and melody pattern) has been modified

before

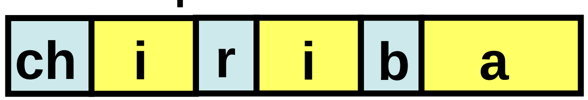


after

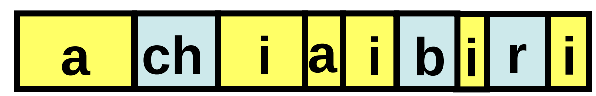


- Prolonged vowels of singing voices has been divided into small pieces

before



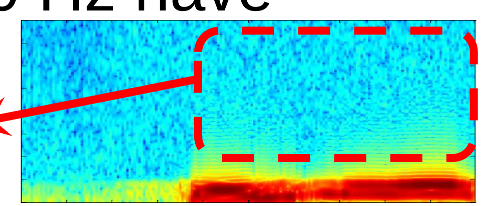
after



Low-pass filtering technique

- Frequency components higher than 800 Hz have been eliminated

Important acoustic cues for discrimination ??



The goal of this study

Subjective experiments

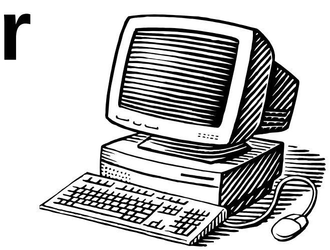
Short-term spectral feature
Temporal structure

Importance !

Based on knowledge obtained by subjective experiments

Automatic vocal style discriminator

- Spectral feature measure
- F0 derivative measure

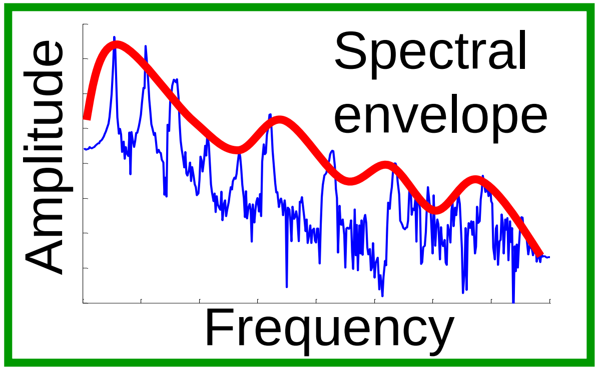


Automatic discrimination measure

- Spectral feature measure

  Difference in spectral envelopes and vowel durations

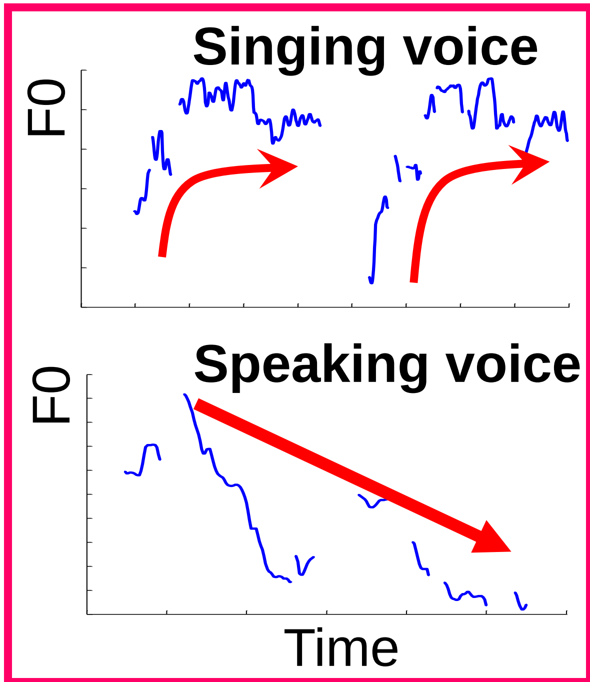
  - Mel-Frequency Cepstrum Coefficients (**MFCC**)
  - **ΔMFCC** (5-frame regression)



- F0 derivative measure

  Difference in dynamics of prosody

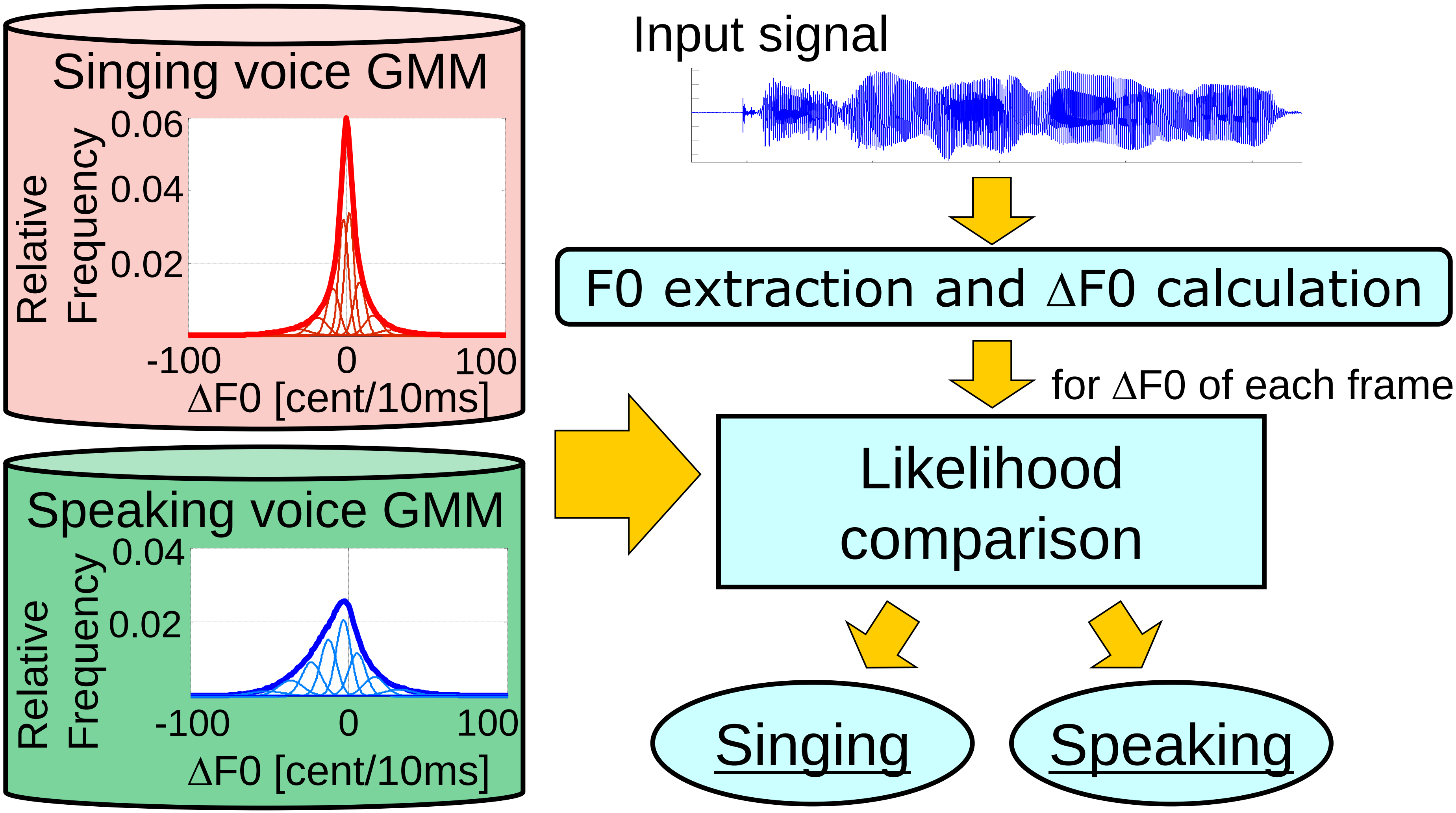
  - **ΔF0** (5-frame regression)
    - F0 Extraction (*PreFEst*, Goto1999)



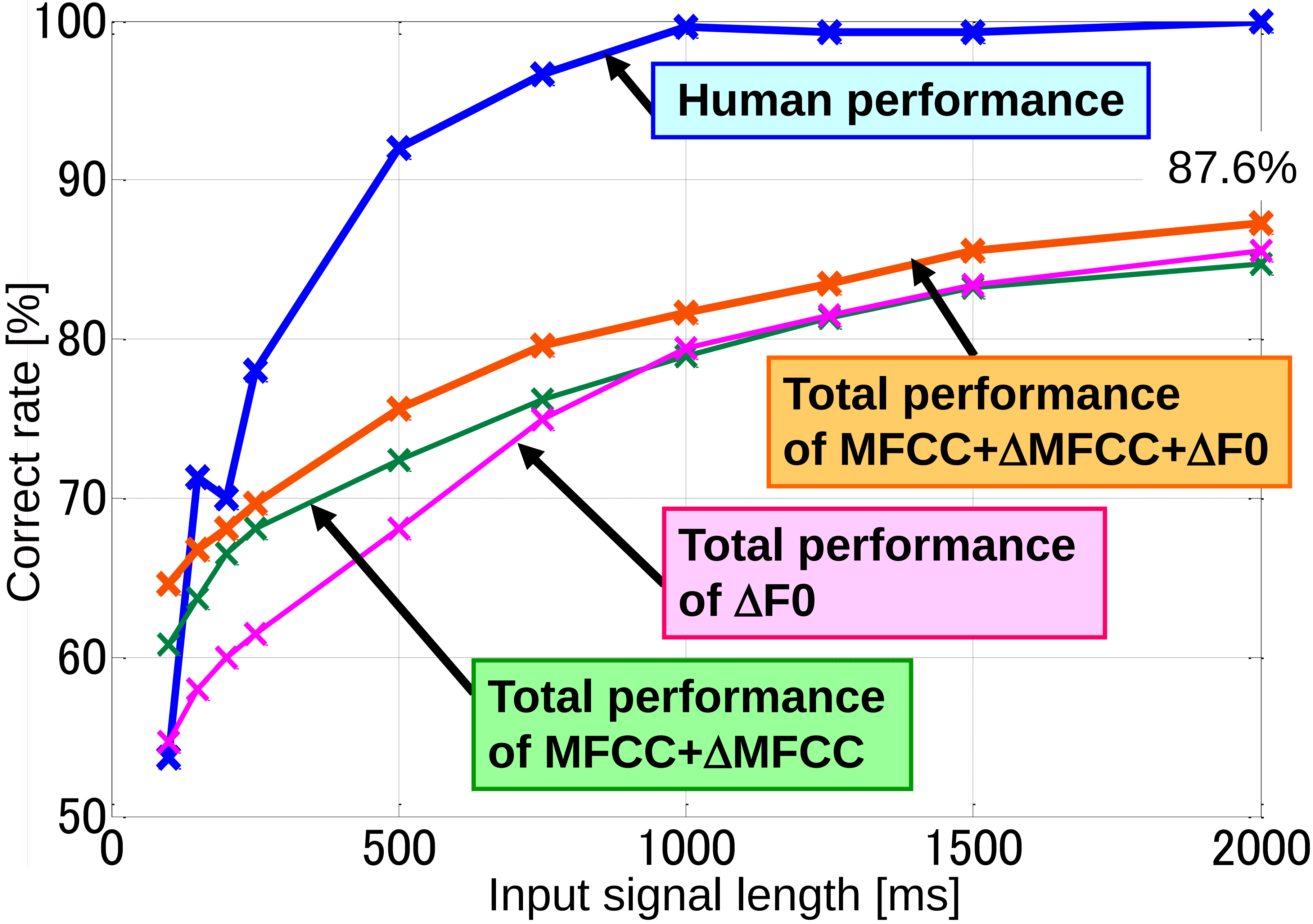
Training the discriminative model

- Gaussian mixture models (16-mixture GMM)

e.g. Discrimination using ΔF0



Automatic discrimination results



Summary and future work

- Investigation of signal length necessary
  - Not only ***temporal characteristics*** but also ***short-time spectral feature*** can be a cue for the discrimination
- Investigation of acoustic cues necessary
  - ***The relative importance of the temporal structure*** is found for singing and speaking voice discrimination
- Automatic vocal style discriminator
  - Feature vector (MFCC+ΔMFCC+ΔF0)
  - For 2-s signals, the correct rate is 87.6%

- Plan to propose ***new measures*** to improve the automatic discrimination performance