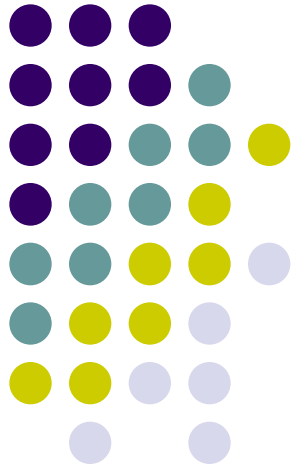
Discrimination between Singing and Speaking Voices

Yasunori Ohishi[1], Masataka Goto[2], Katsunobu Itou[1] and Kazuya Takeda[1]

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

[2]National Institute of Advanced Industrial Science and Technology



Introduction

- Automatic discrimination system of voices

Discrimination
between **Singing** and **Speaking** voices

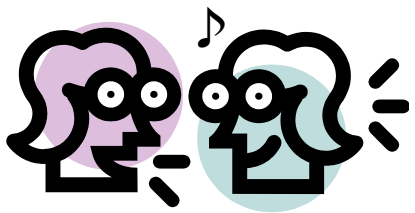
Typical characteristics of the singing voice

- F0 and intensity vary widely
- *Singing Formant*

⇒ Not necessarily heard in an amateur's singing voice

⇔ Possible for humans to discriminate just by listening to an amateur's singing voice ♪ ♪

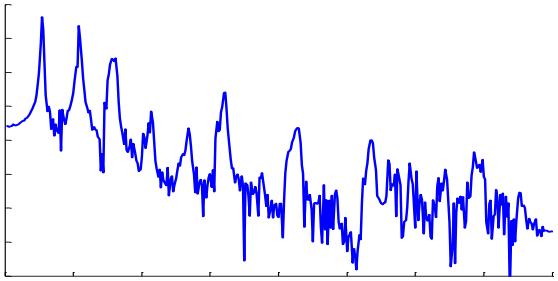
What features should we extract?



Early research

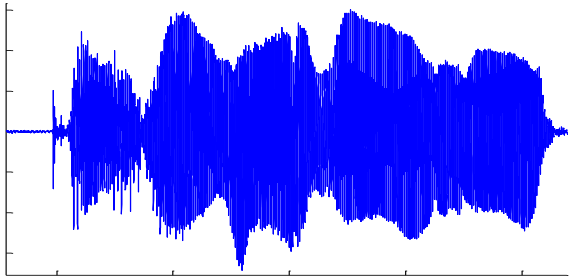
- Algorithms that discriminate “music” from “speech”

Frequency domain



Spectral Centroid, MFCC, etc…

Time domain



Zero Cross Rate, etc…

“music” category

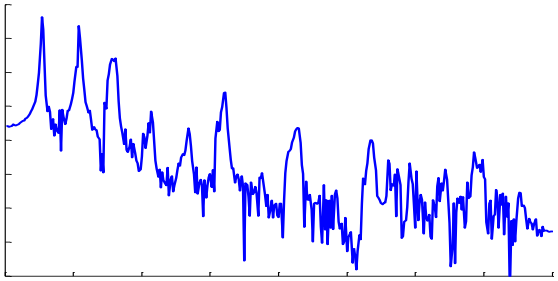
- Instrumental sounds and the singing voice with accompanying sounds

  The characteristics of **the singing voice without accompaniments** yet to be fully discussed

Early research

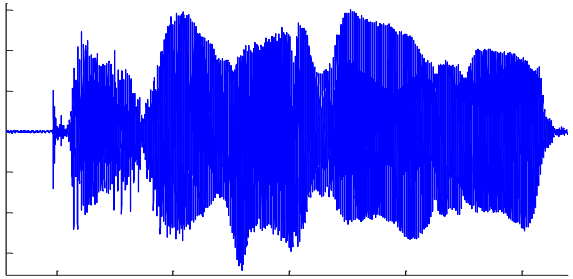
- Algorithms that discriminate “music” from “speech”

Frequency domain



Spectral Centroid, MFCC, etc…

Time domain



Zero Cross Rate, etc…

The goal of this study is to

Characterize the nature of the singing voice

Build a measure to discriminate the singing voice from the speaking voice

Quiz!!

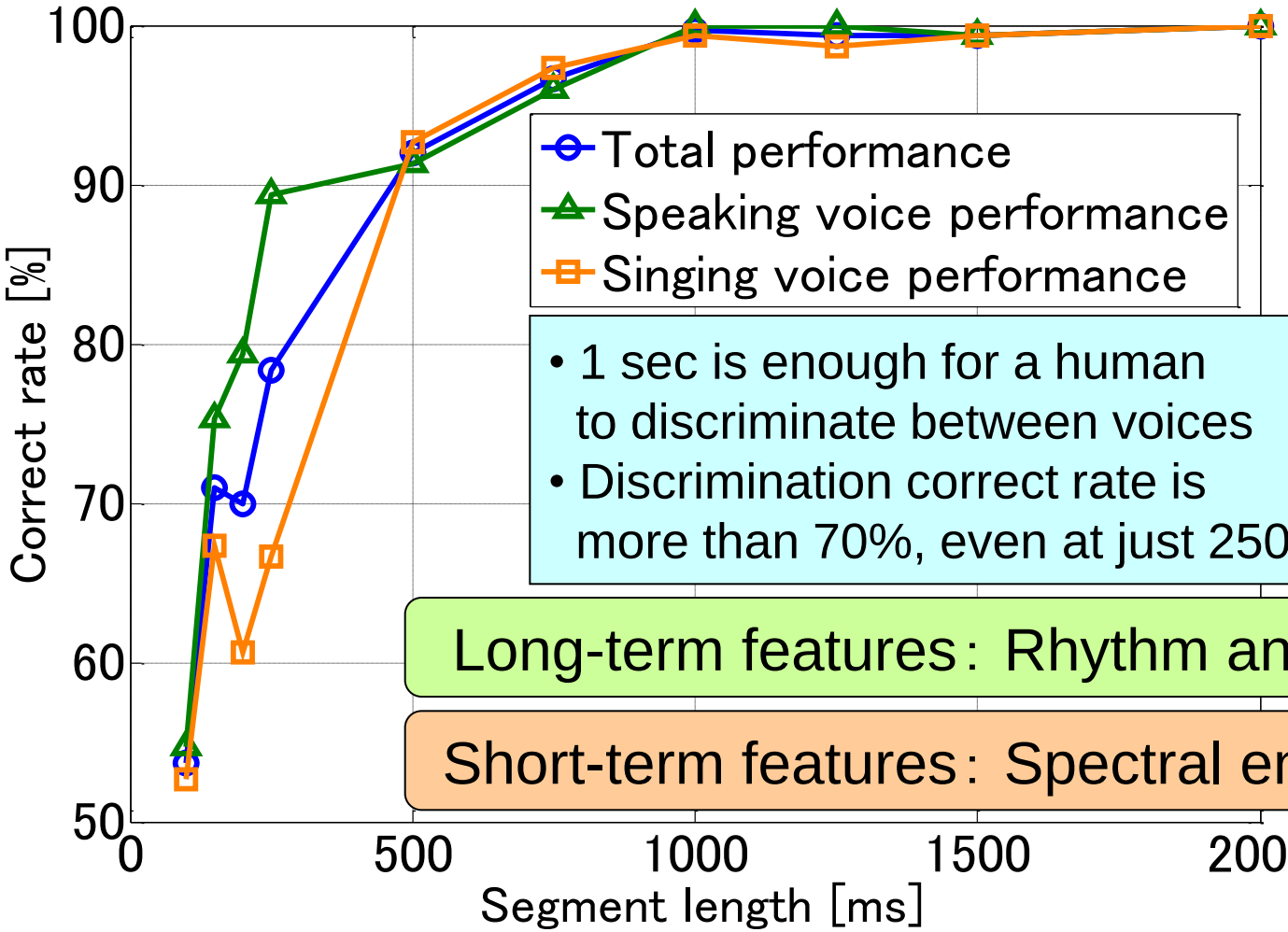
- Can you discriminate between **Singing** and **Speaking** voices?

(Japanese voices)

Q1. Can you do it ? (2s long)	
Q2. Can you do it ? (500ms long)	
Q3. Can you do it ? (250ms long)	

Results of human performance

- 10 Japanese subjects answered whether the voice was singing or speaking, or that it was impossible to identify them

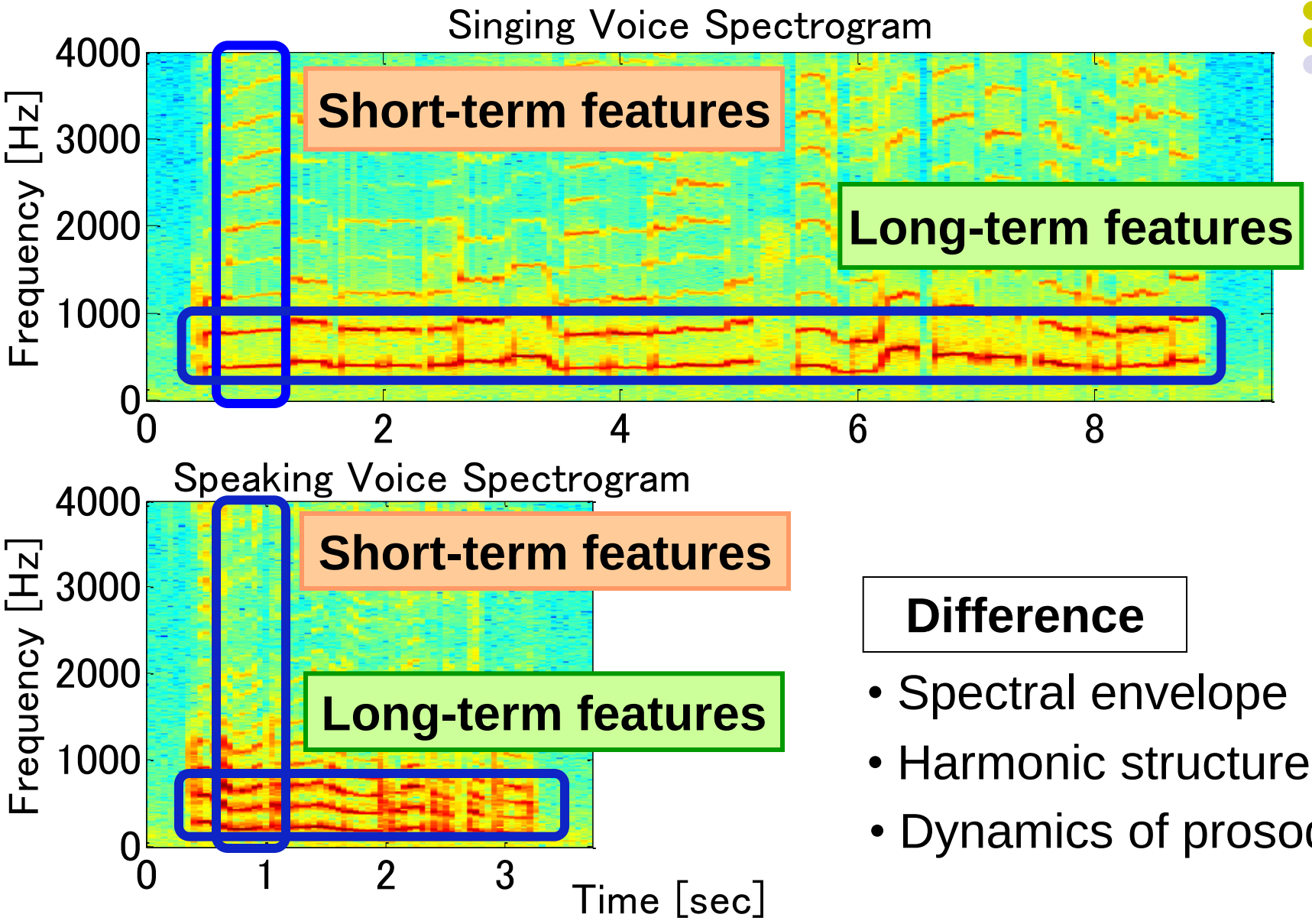


- 1 sec is enough for a human to discriminate between voices
- Discrimination correct rate is more than 70%, even at just 250 ms

Long-term features: Rhythm and Melody

Short-term features: Spectral envelopes

Discrimination Measures



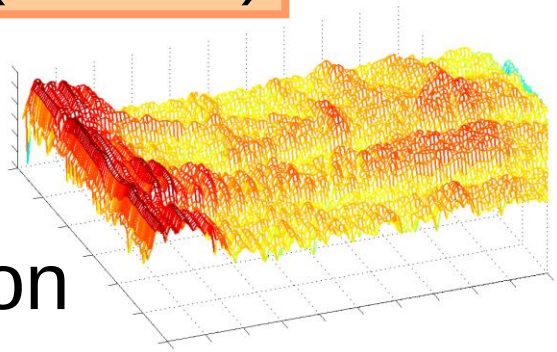
Difference

- Spectral envelope
- Harmonic structure
- Dynamics of prosody

Short-term feature measure

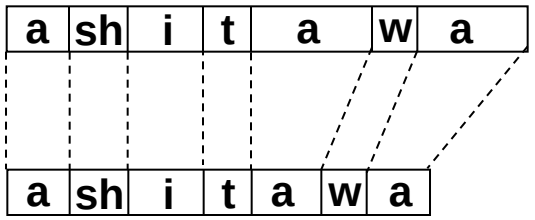
- The difference in the spectral envelopes

Mel-Frequency Cepstrum Coefficients (MFCCs)



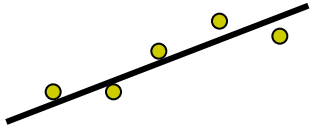
  - Calculating for a 100-ms hamming windowed frame every 10 ms

- The difference in the vowel duration
  - As for the singing voice, prolonged phonemes can be observed frequently
  - The phoneme in the speaking voice changes one after another



ΔMFCCs (MFCC derivatives)

  - Regression parameters over 5 frames

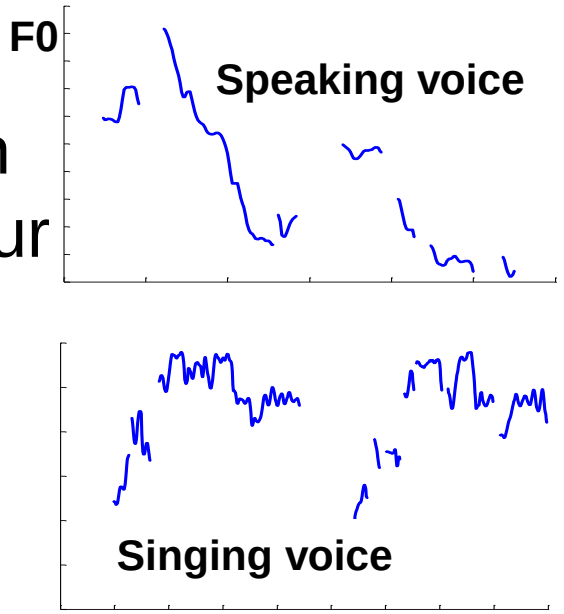


Long-term feature measure

- F0 Extraction
  - A predominant-F0 estimation method (*PreFEst*)
  - Calculating the F0 value for every 10 ms
  - Smoothing by a median filter
- The difference in the dynamics of prosody

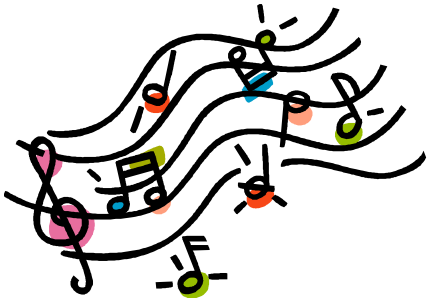
ΔF0 (five-point regression)

  - Japanese speaking voice's intonation is characterized by a falling F0 contour
  - Singing voice is generated under the constraint of melodic and rhythm patterns



Introducing the voice database

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



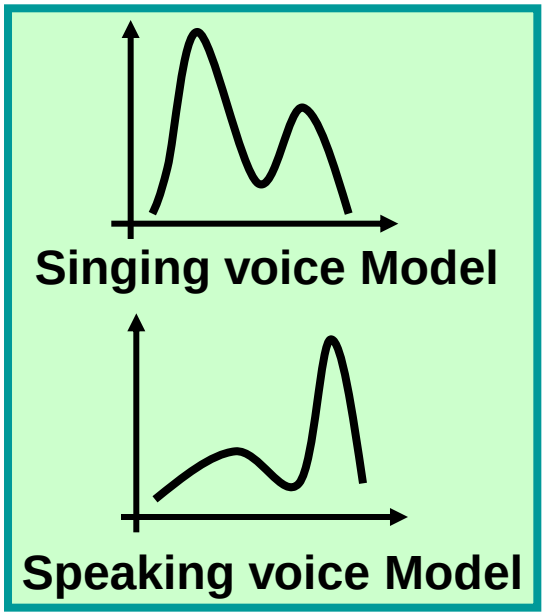
  - Reading the lyrics of chorus and “verse A” sections

  - A total of 100 samples per subject
    (Singing voice: 50 samples, Speaking voice: 50 samples)

Evaluation of the Proposed Method

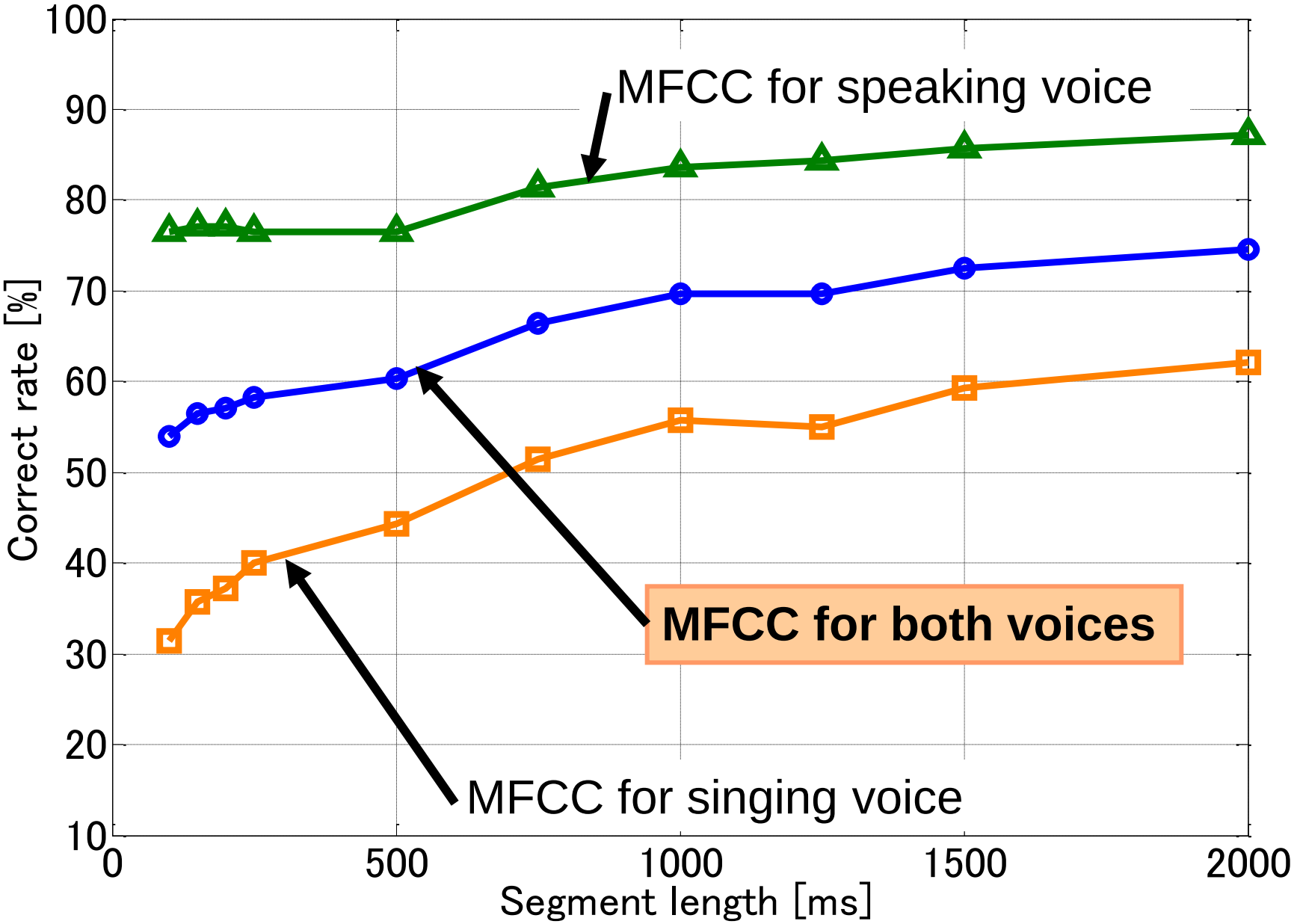
- Training the 16-mixture GMMs
  - 2,500 sound signals of the singing and speaking voices of 25 subjects
- Testing the proposed method
  - 480 sound samples from 50 subjects
  - The maximum likelihood principle

$$\hat{d} = \underset{d=\text{sing},\,\text{speak}}{\arg\max}\ f(\mathbf{x};\Lambda_d)$$

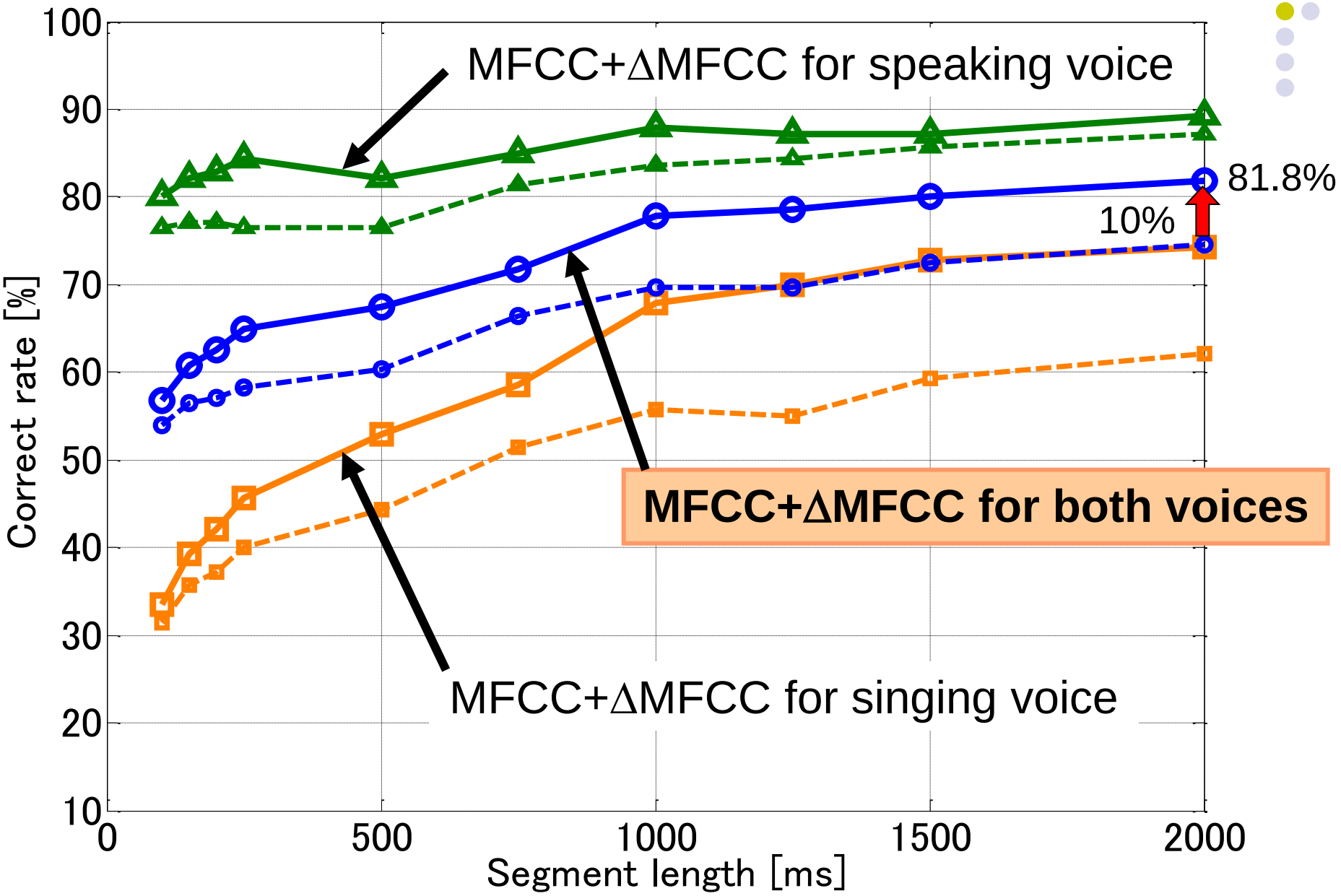


- $\Lambda_d, (d = \text{sing}, \text{speak})$ are the GMM parameters
  (Weight, Mean, Variance)

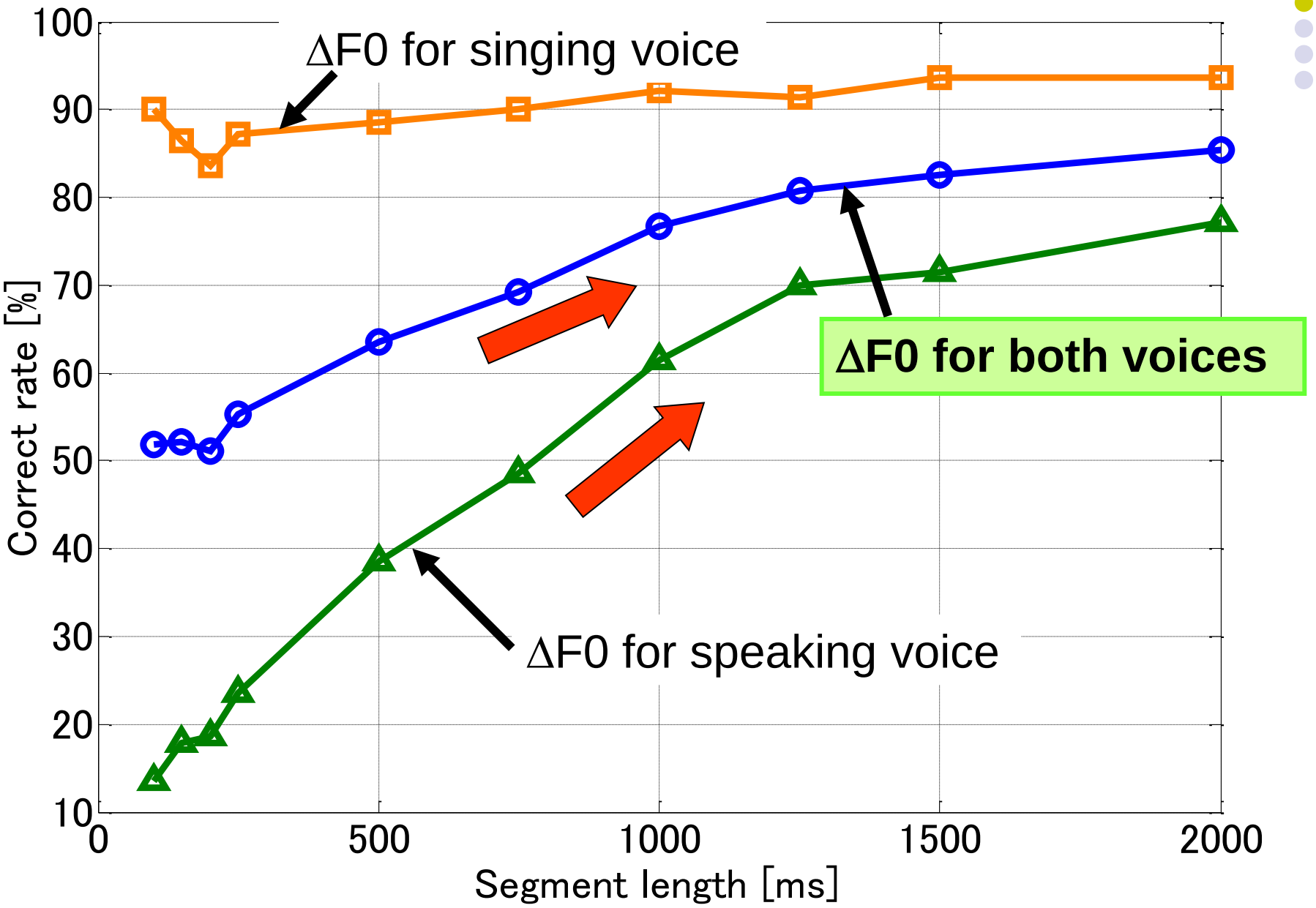
Performance of short-term features



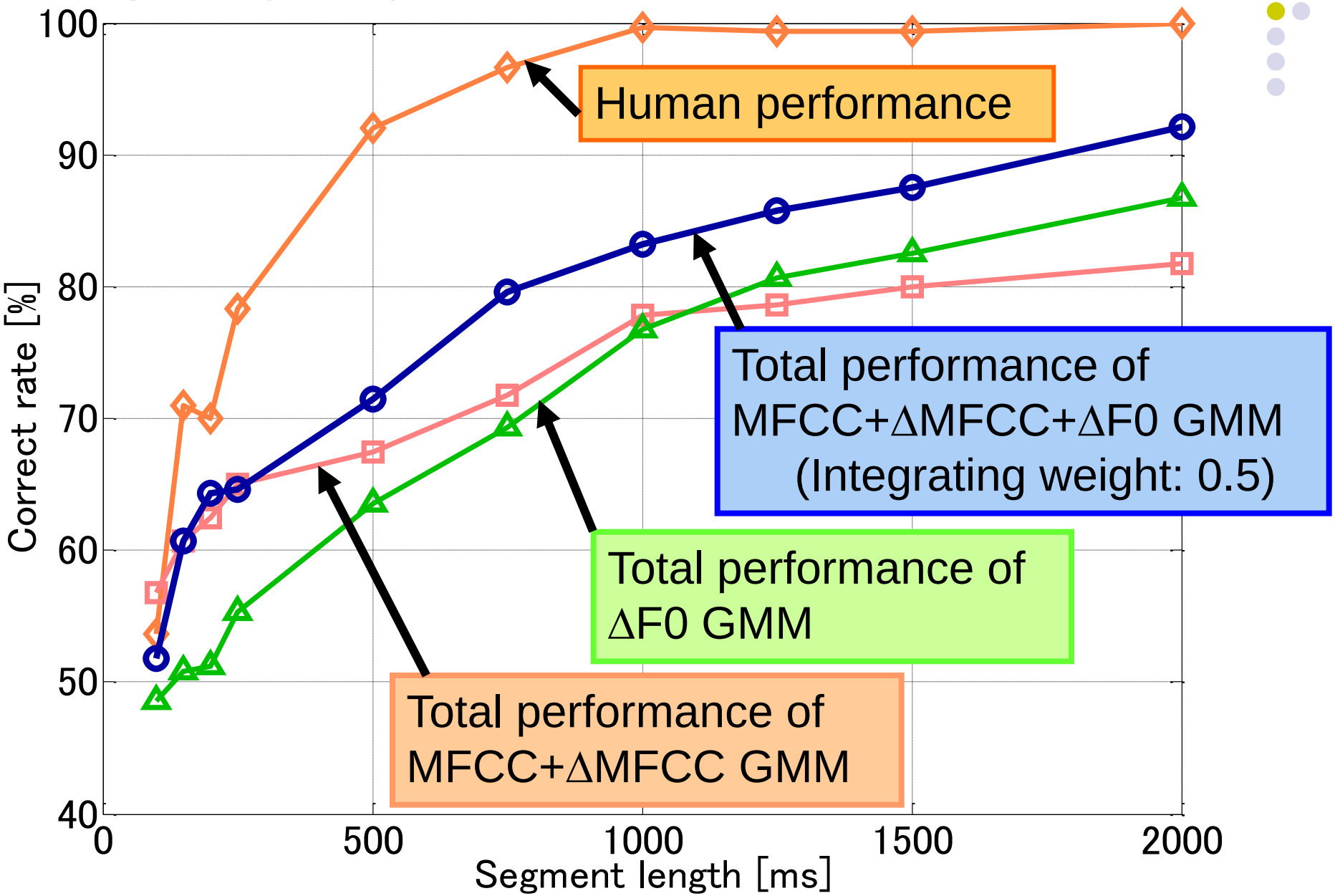
Performance of short-term features



Performance of long-term feature

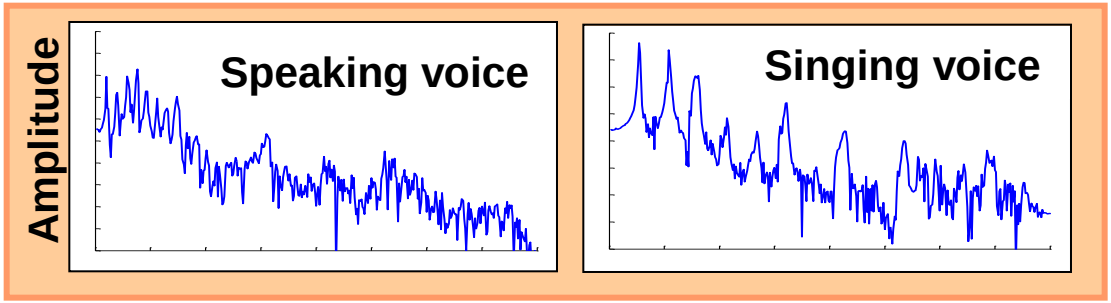


Comparing long-term and short-term features

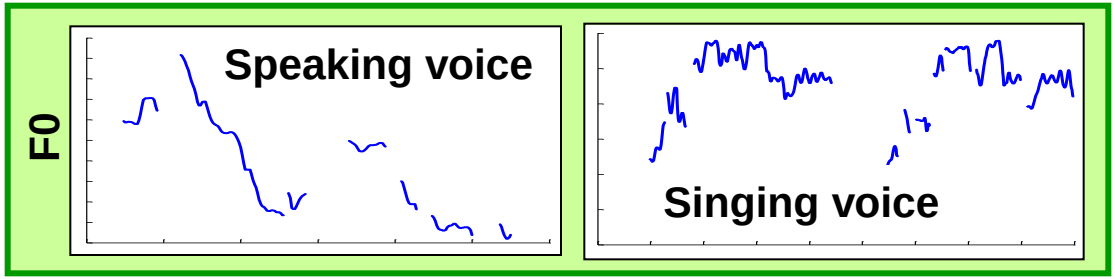


Discussion

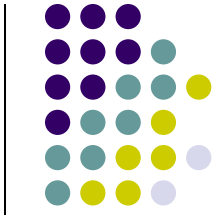
- Capturing the signal features **effectively** and **complementarily**
  - The difference between the spectrum envelopes are dominant cues for the discrimination of signals shorter than one second



  - ΔF0 is effective for the signals of one second or longer by dealing with the difference of global F0 contours



Summary and Future work



- Discrimination of singing and speaking voices by modeling two different aspects
  - 65.0% correct rate even with 250-ms signals by MFCC
  - 85.4% correct rate with 2-sec signals by ΔF0
- Simple combination of the two measures
  - 92.1% correct rate with 2-sec signals

- To optimize the integration method for the two complementary measures
- To extract the time domain features

Music Retrieval for automatic Voice Discrimination System