歌声と朗読音声の識別システム構築のための人間の識別能力の調査

大石 康智[1], 後藤 真孝[2]
伊藤 克亘[1], 武田 一哉[1]
[1]名古屋大学大学院情報科学研究科
[2]産業技術総合研究所

はじめに

- 歌声と朗読音声の自動識別手法の提案
  - 歌声とその歌詞を朗読した音声の識別

歌声 

朗読音声 

特徴量

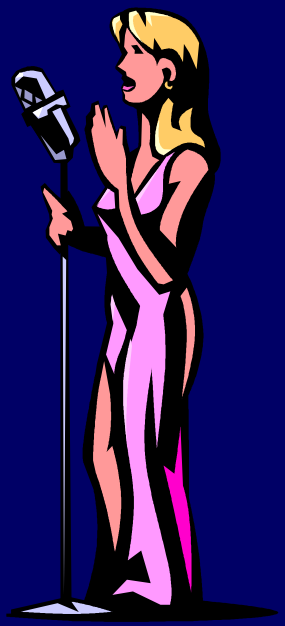
- 音色の違い　　MFCC
- 音高の変化の違い　　ΔF0

人間は,
- どの程度の音声信号長があれば識別可能か?
- どのような特徴を手がかりとして識別を行うのか?

さてここで問題です

みなさんは歌声と朗読音声の識別ができますか?

問題1	1秒の音声	
問題2	500msの音声	
問題3	250msの音声	

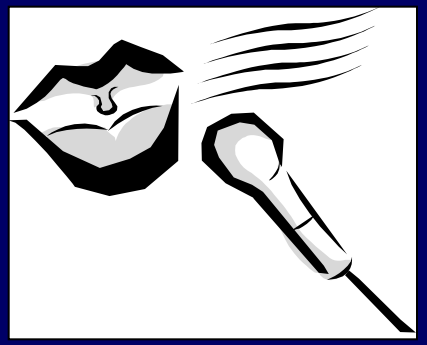


歌声データベース

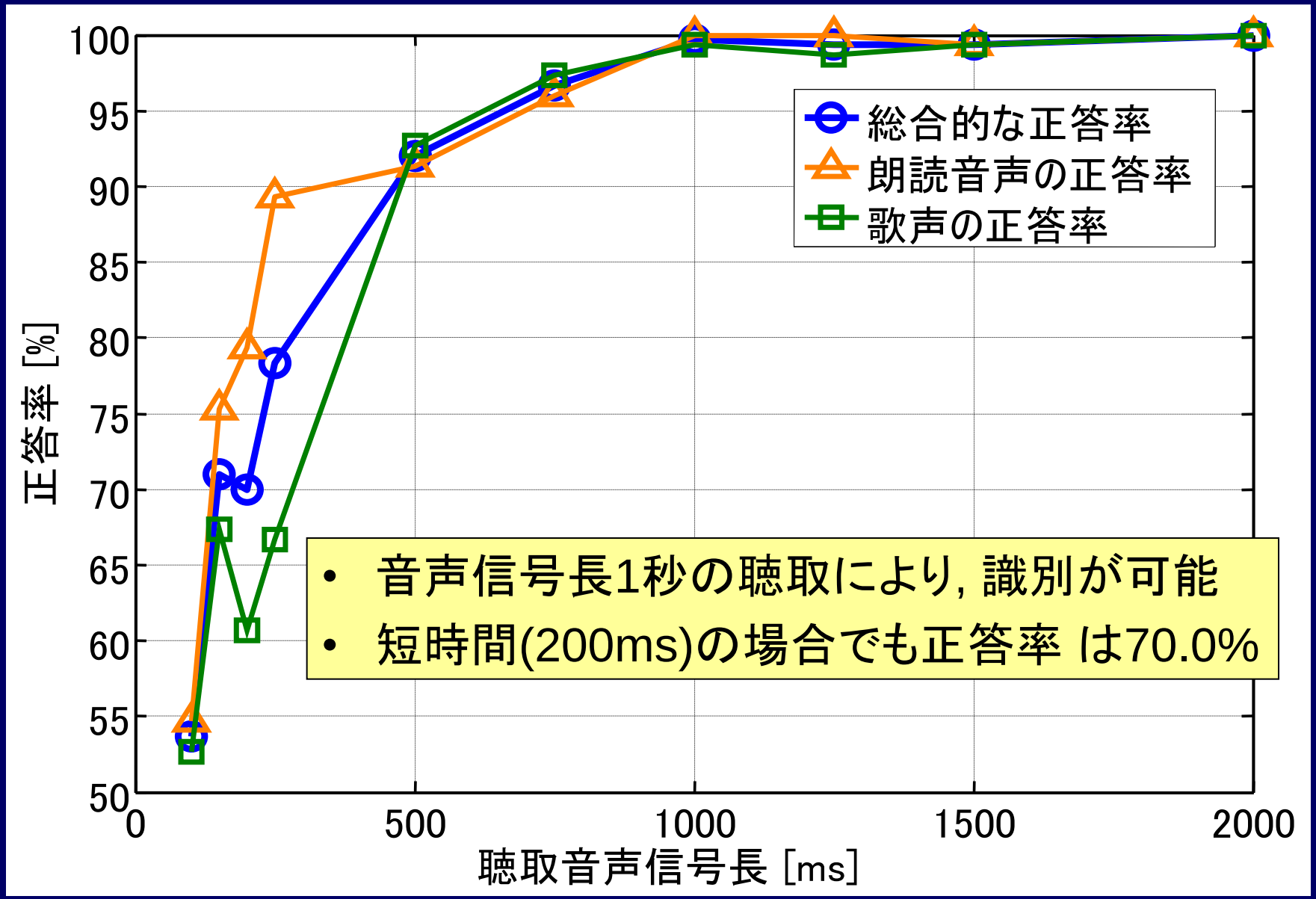
- AISTハミングデータベース
  (歌声研究用音楽データベース)
  - 被験者がある曲の出だしとサビの部分を歌う
  - またその歌詞を朗読する

収録被験者1名あたり計100サンプル（日本人75名)
(歌声: 25曲 x 2パート, 朗読音声: 25曲 x 2パート)

  - 収録音声サンプルの長さ
    - 歌声:　　平均12.0秒
    - 朗読音声: 平均7.0秒



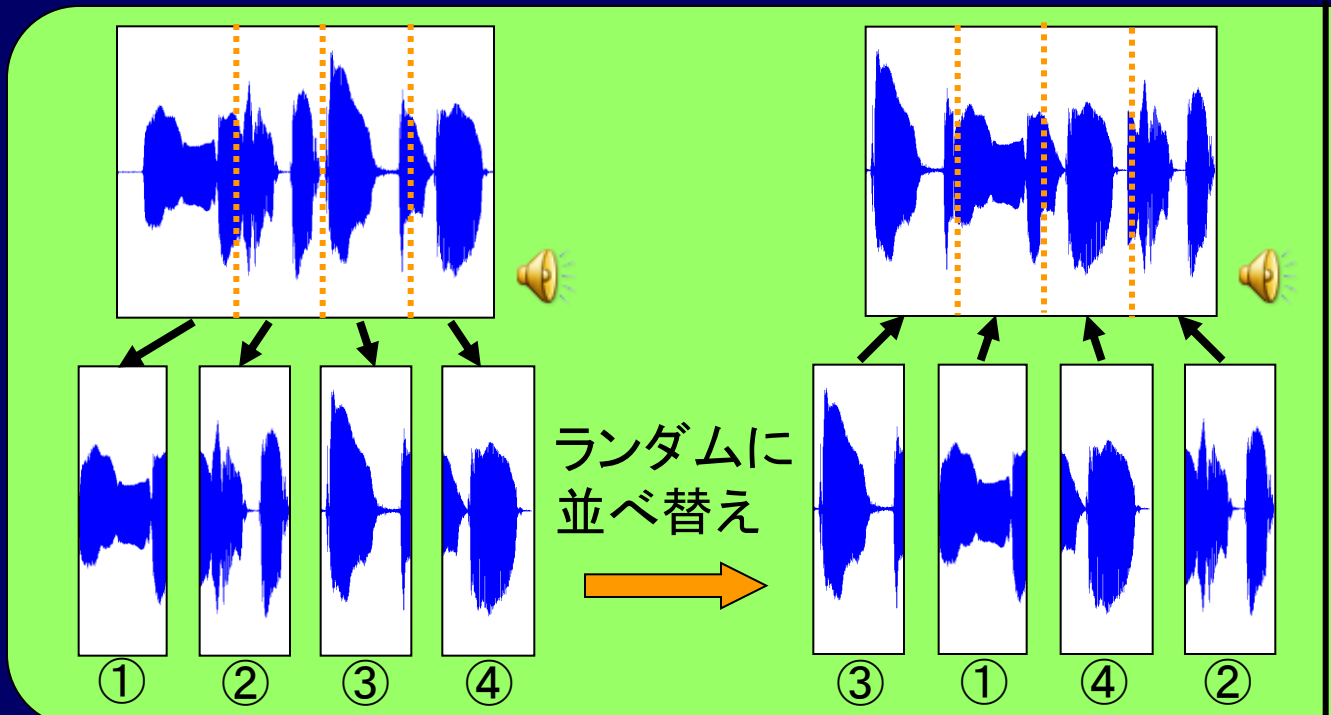
識別に必要な音声信号長の調査



識別に影響する音声信号の特徴の調査

- Random Splicing手法
  - 音声サンプルをある長さの断片に分割し，ランダムに接合する

➡ メロディのパターン，テンポ，リズムをマスク



音声サンプル
(1秒)

分割長250ms

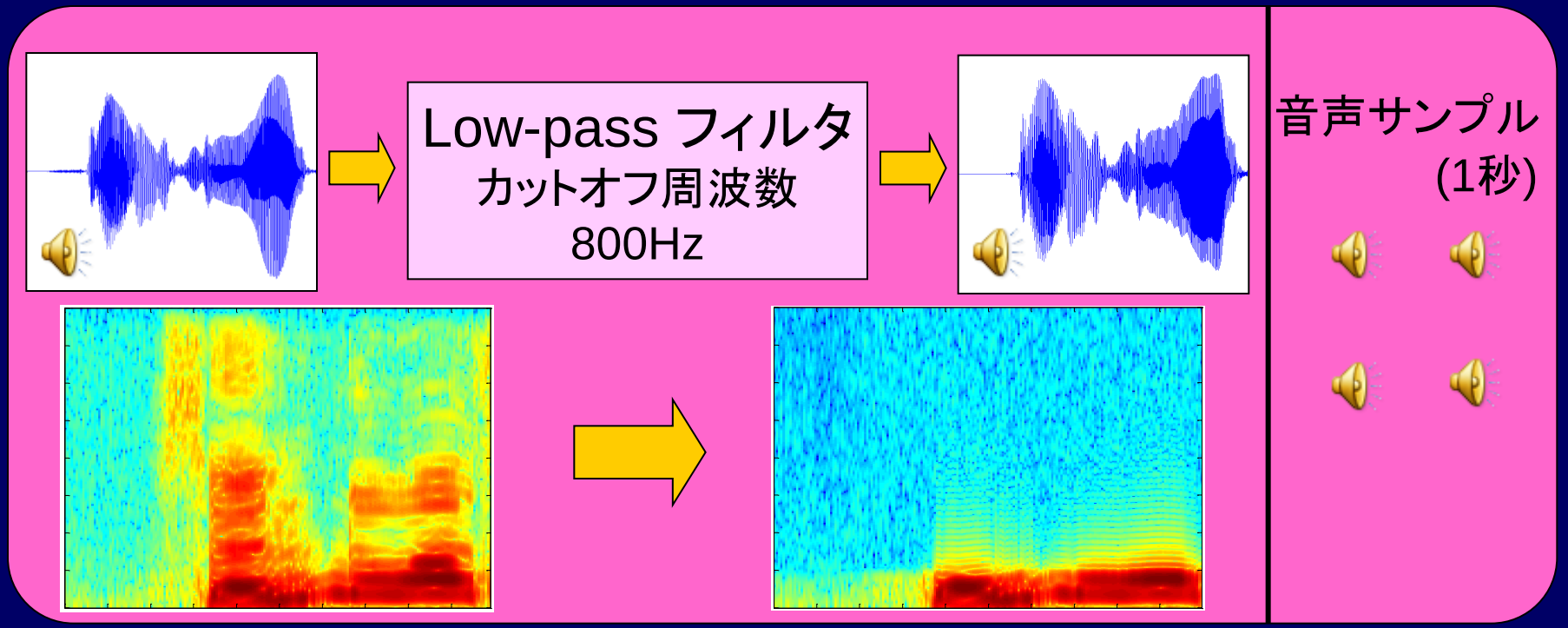
分割長200ms

分割長125ms

識別に影響する音声信号の特徴の調査

- Filtering手法
  - ローパスフィルタにより
    音声信号の高調波成分を除去

➡ 音色，音質の低下



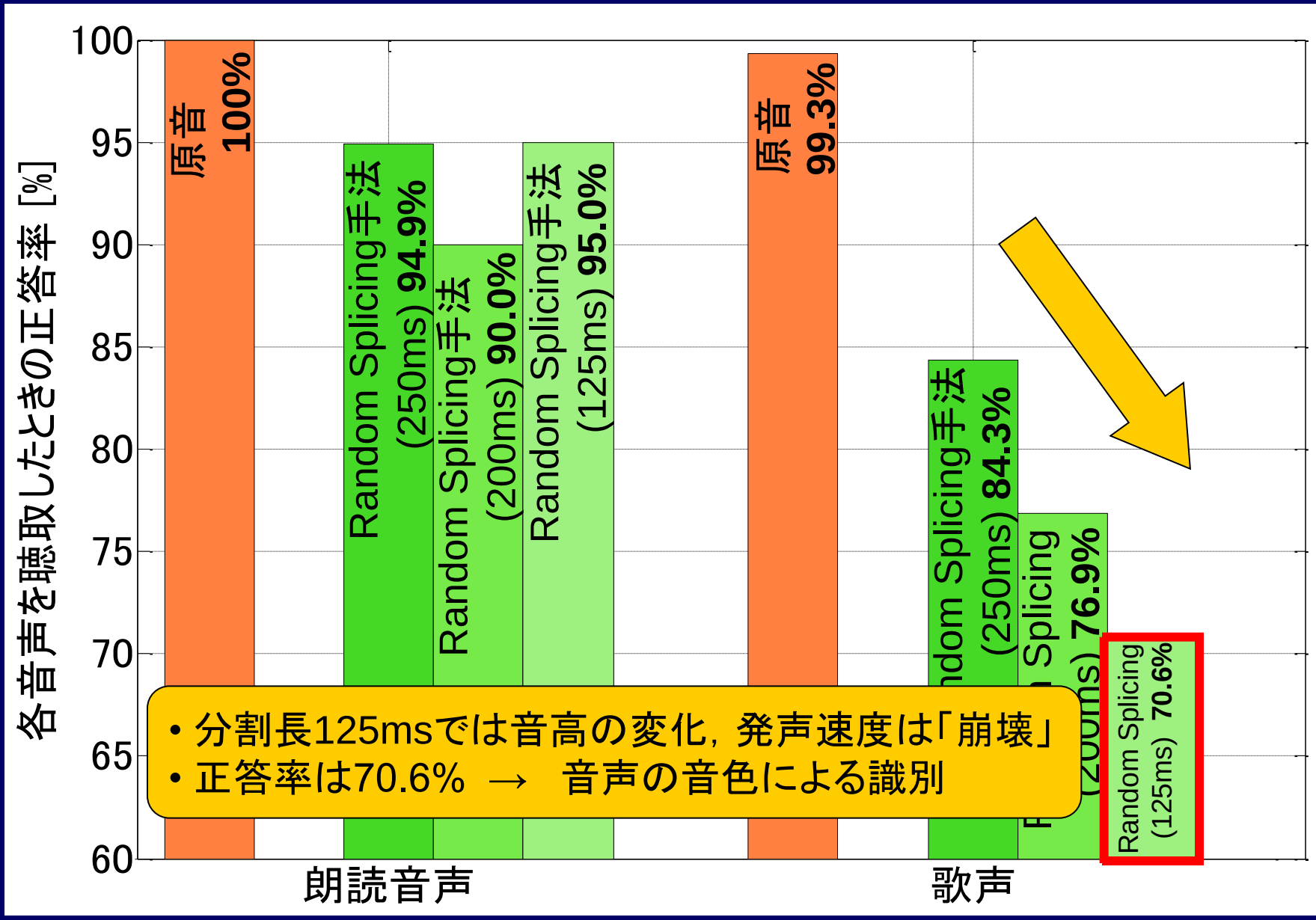
聴取実験

- 識別に必要な音声信号長の調査より

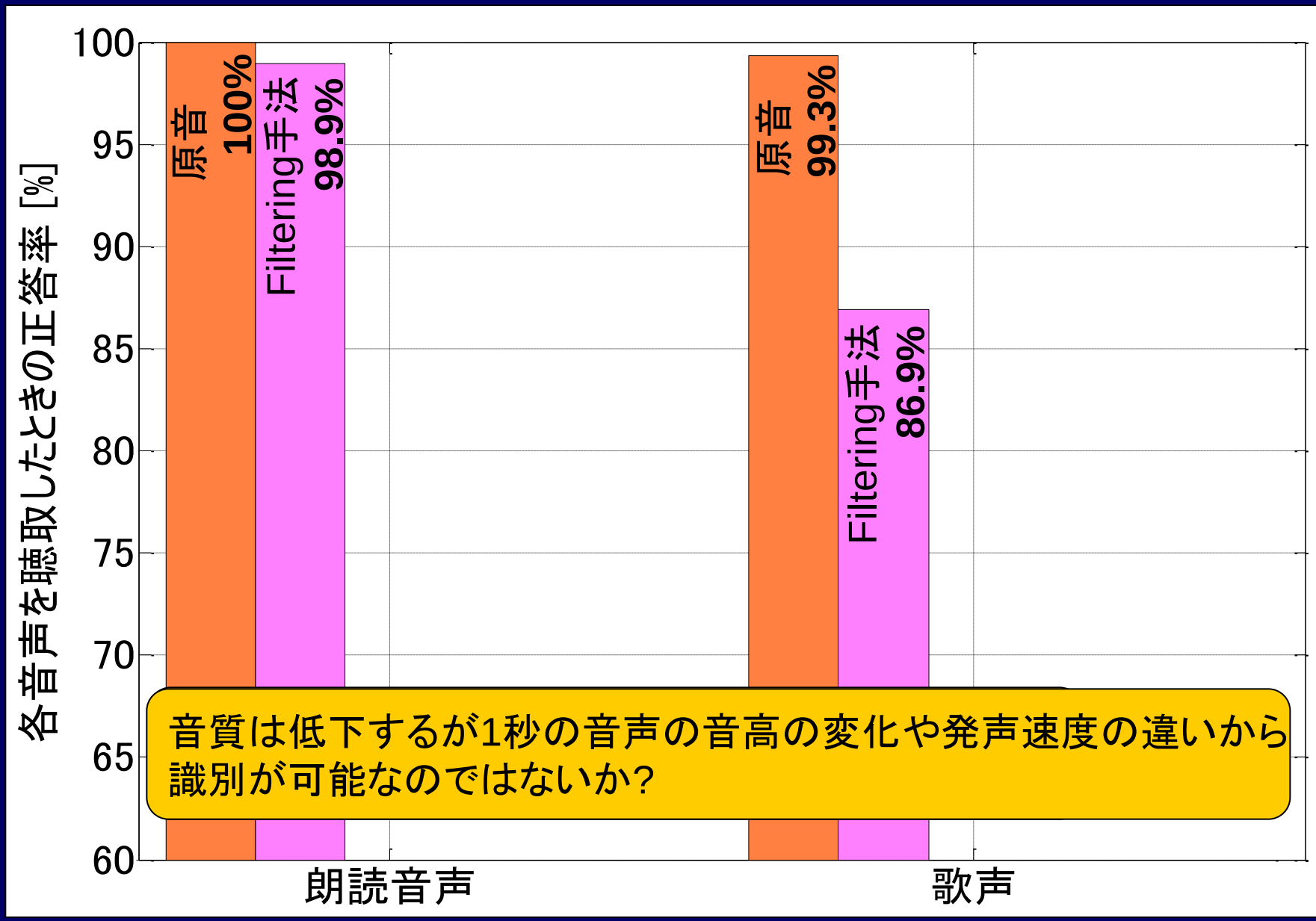
  ➡ 1秒の音声信号長があれば約100%識別可

  1秒の音声信号に対して
  - Random Splicing手法
    - 分割長(250ms，200ms，125ms)
  - Filtering手法
- 被験者　10名
  - 聴取した音声が歌声か朗読音声か?

加工音声の聴取実験結果

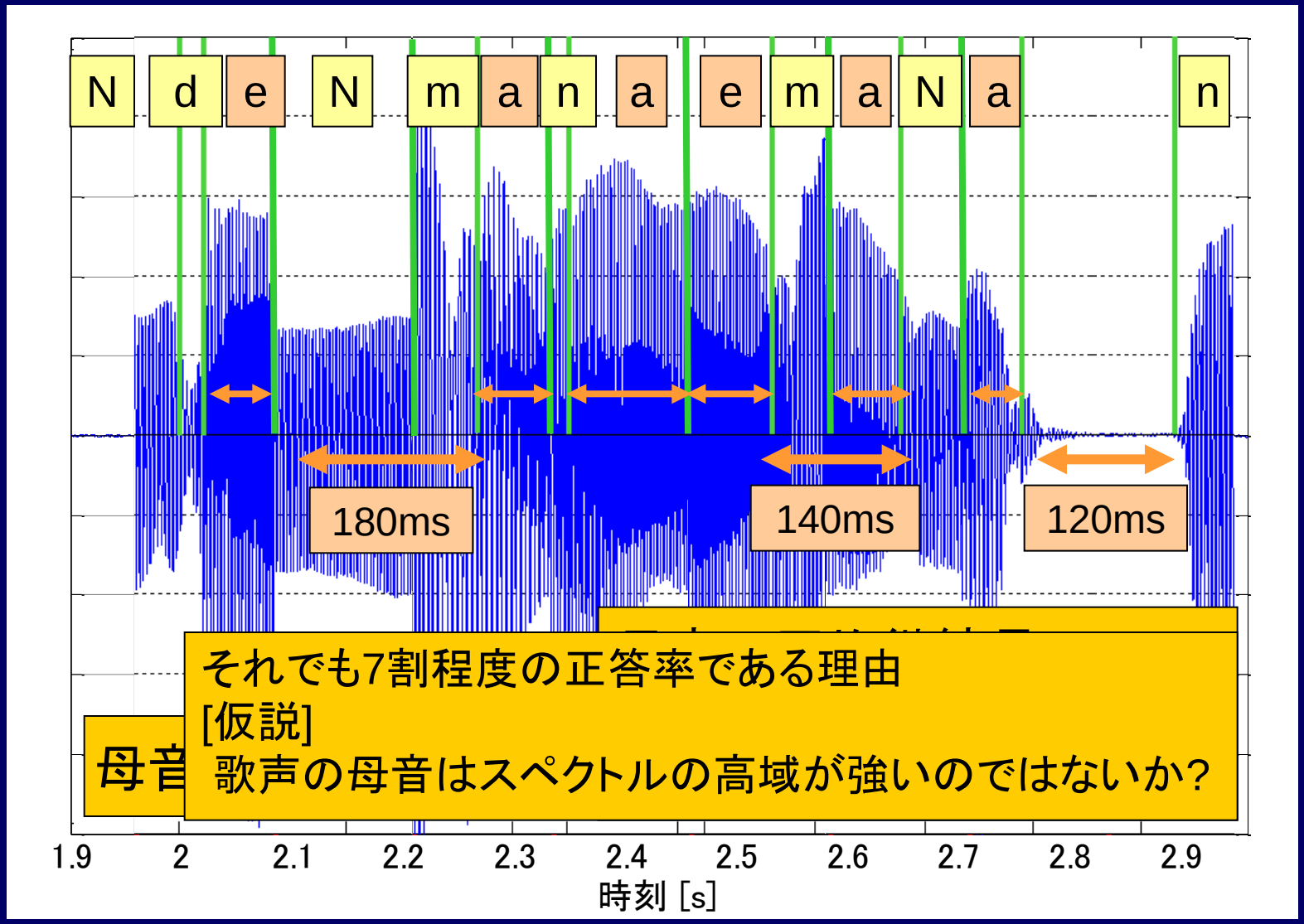


加工音声の聴取実験結果



歌声の正答率低下に対する考察

- 音素(母音)の継続長の変化



自動音声識別手法との比較

- 特徴量

  音声の音色の違い:　**MFCC**(12次)+**ΔMFCC**(12次)

  音高の変化の違い:　**ΔF0**

  (Δ算出は50msの窓幅)

- 識別方法

  16混合ガウス分布の事後確率による識別

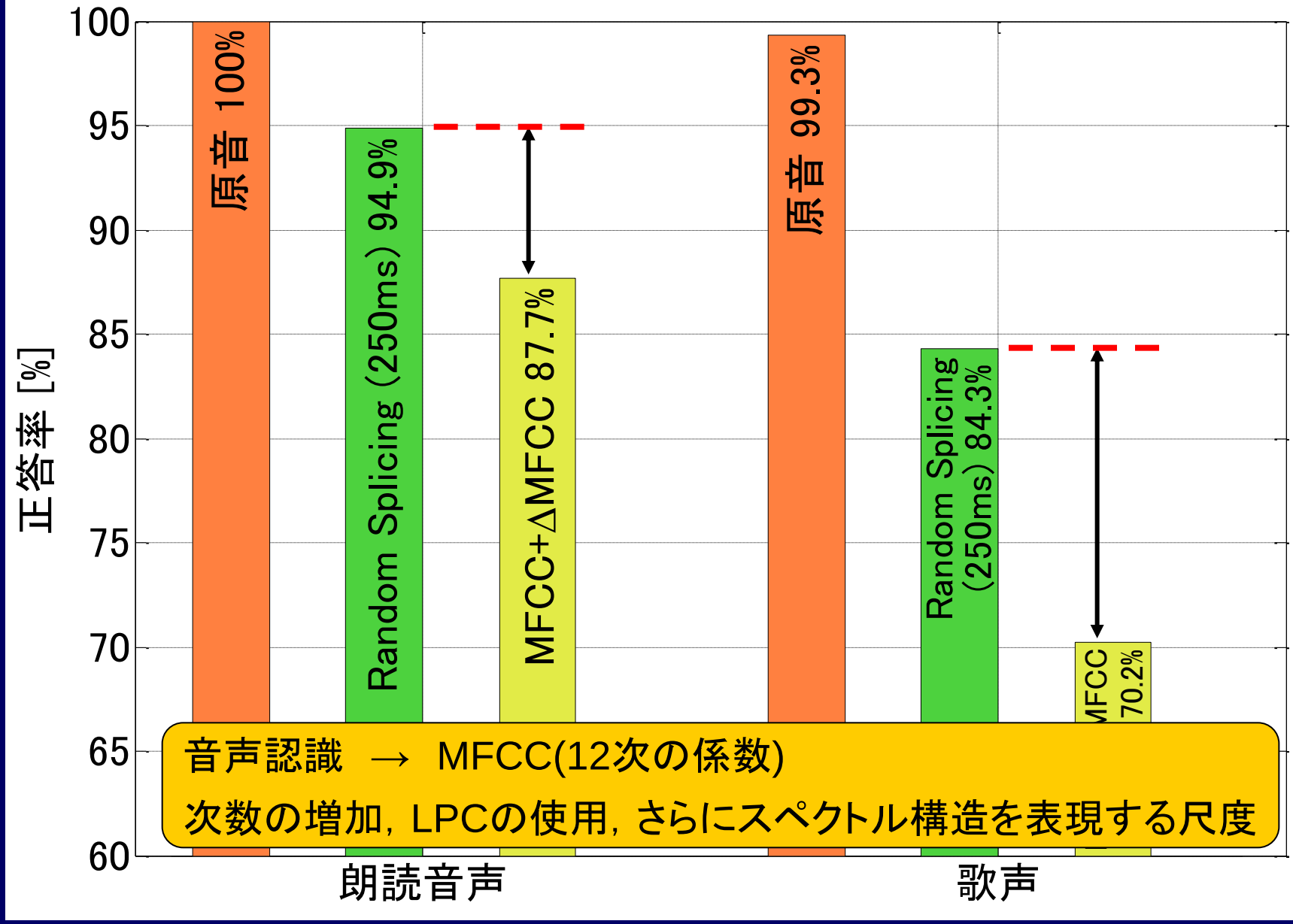
音声の音色の違いによる識別

Random Splicing手法 ⟺ **MFCC+ΔMFCC**

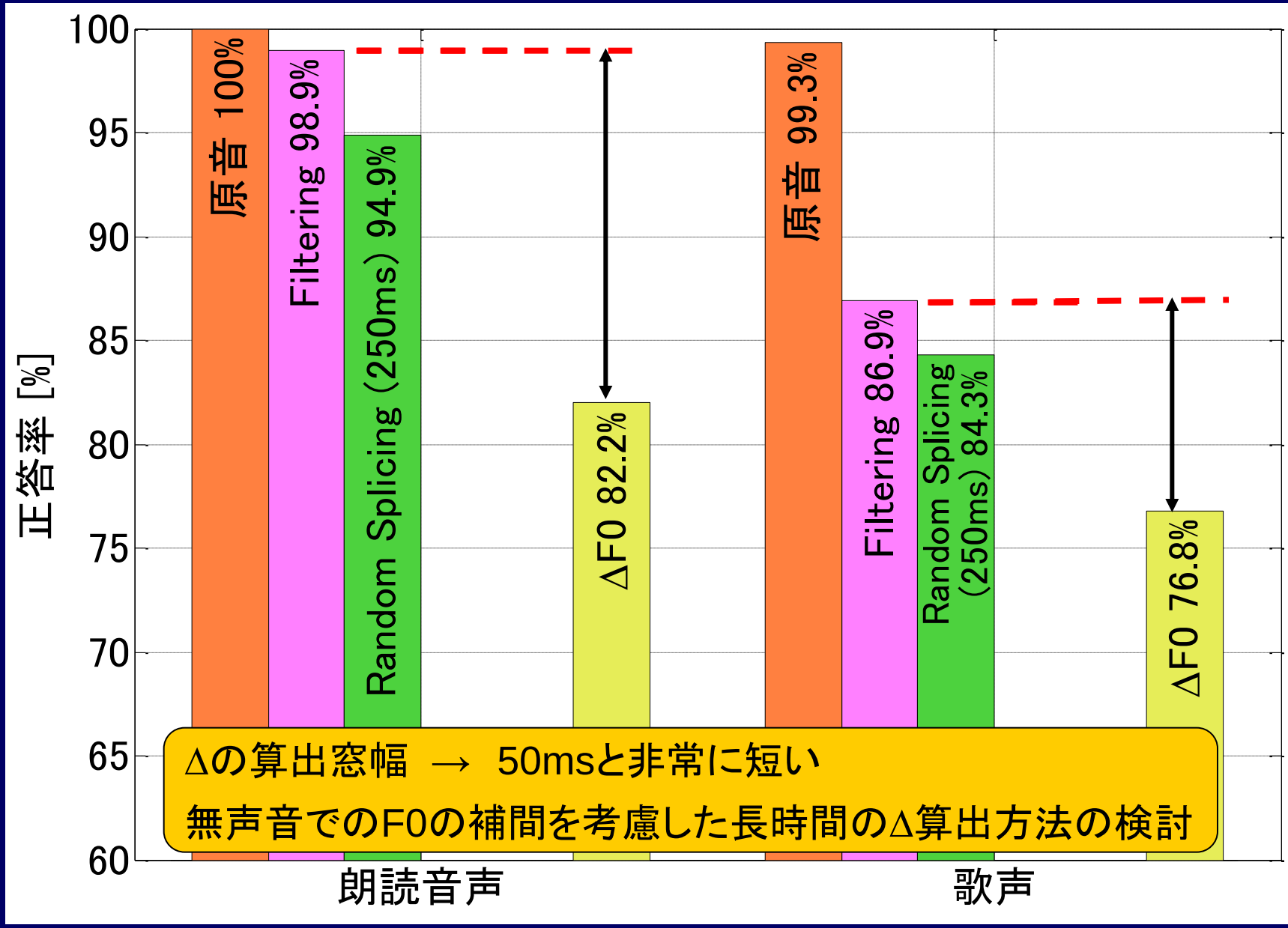
音高の変化の違いによる識別

Filtering手法 ⟺ **ΔF0**

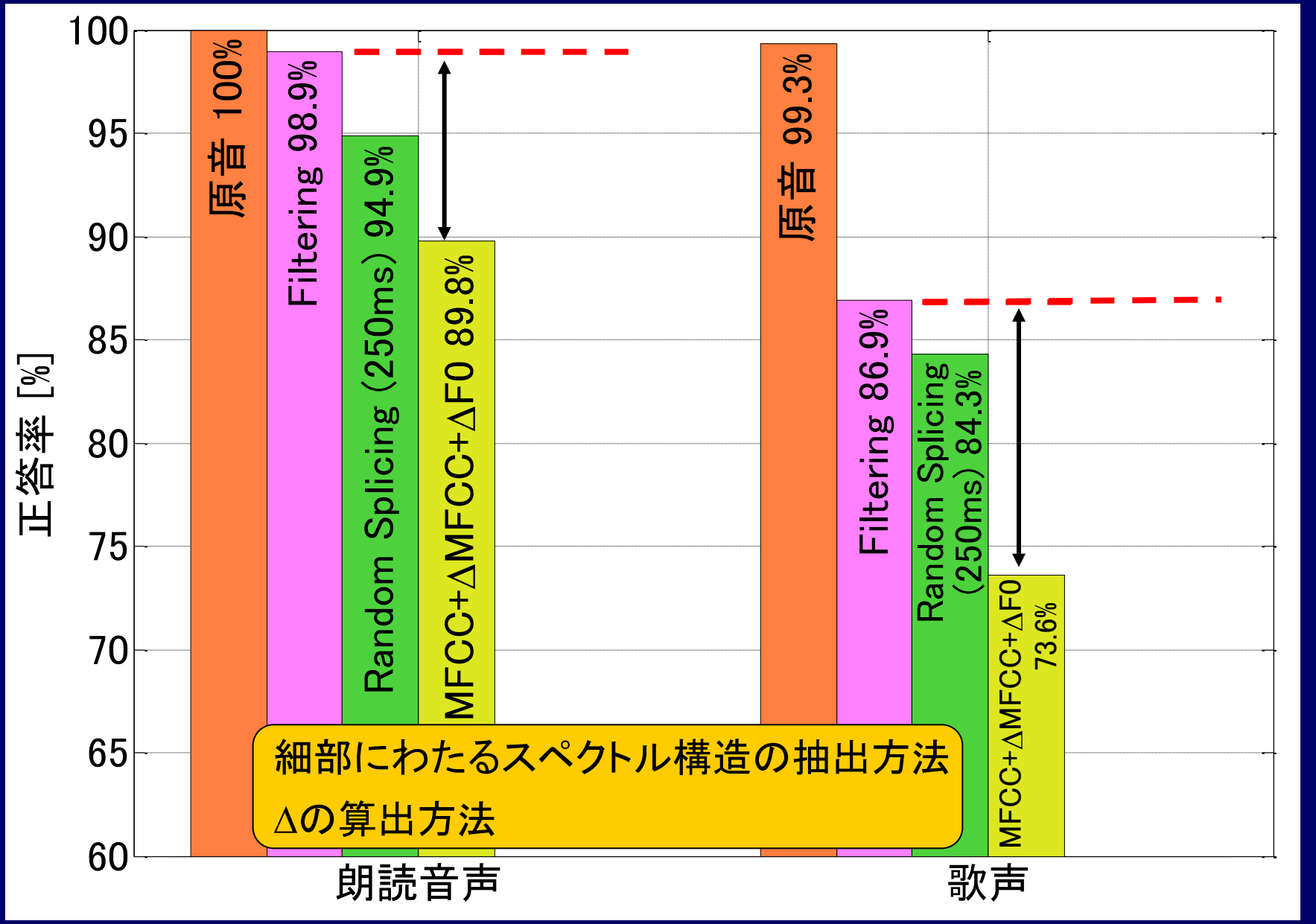
自動音声識別手法との比較



自動音声識別手法との比較



自動音声識別手法との比較



まとめ

- 人間による音声信号の識別能力の調査
  - 識別に必要な音声信号長の調査
    - 250msの音声信号: 78.3%
    - 1sの音声信号: 99.7%
  - 識別に影響する音声信号の特徴の調査
    - Random Splicing手法
    - Filtering手法
    - 歌声の正答率の低下
  - 聴取結果とシステムの性能との比較
    - 聴取能力と自動識別手法の正答率の差は20%

今後の展開

- 聴取実験で誤識別されたサンプルの解析
- 特徴量の改善
  - 細部にわたるスペクトル構造の抽出方法
  - 無声音を考慮した長時間におけるΔの算出方法

歌声データベース

- AISTハミングデータベース
  (歌声研究用音楽データベース)
  - 日本人歌唱者75名分（男性37名, 女性38名）
  - ‘RWC Music Database: Popular Music’から抜粋した合計25曲の
  - 歌の出だしの部分とサビの部分を歌う,
    またその歌詞を朗読
  - 1名あたり計100サンプル
    (歌声: 25曲 x 2パート, 朗読音声: 25曲 x 2パート)
  - 音声サンプルの長さは歌声で約8秒, 朗読音声で約5秒

識別に必要な音声信号長の調査

- 評価セットの構成

時間長	歌声	朗読音声
100, 150, 200, 250, 500, 750, 1000ms	25サンプル	25サンプル
1250ms	20サンプル	20サンプル
1500, 2000ms	10サンプル	10サンプル
合計	215サンプル	215サンプル

識別に必要な音声信号長の特徴の調査

- 加工した音声の評価セットの構成

Random Splicing 手法		
分割する長さ	歌声	朗読音声
125ms	40サンプル	40サンプル
200ms	40サンプル	40サンプル
250ms	20サンプル	20サンプル
合計	100サンプル	100サンプル

Filtering 手法		
	歌声	朗読音声
合計	100サンプル	100サンプル

Random Splicingした音声に対する感想

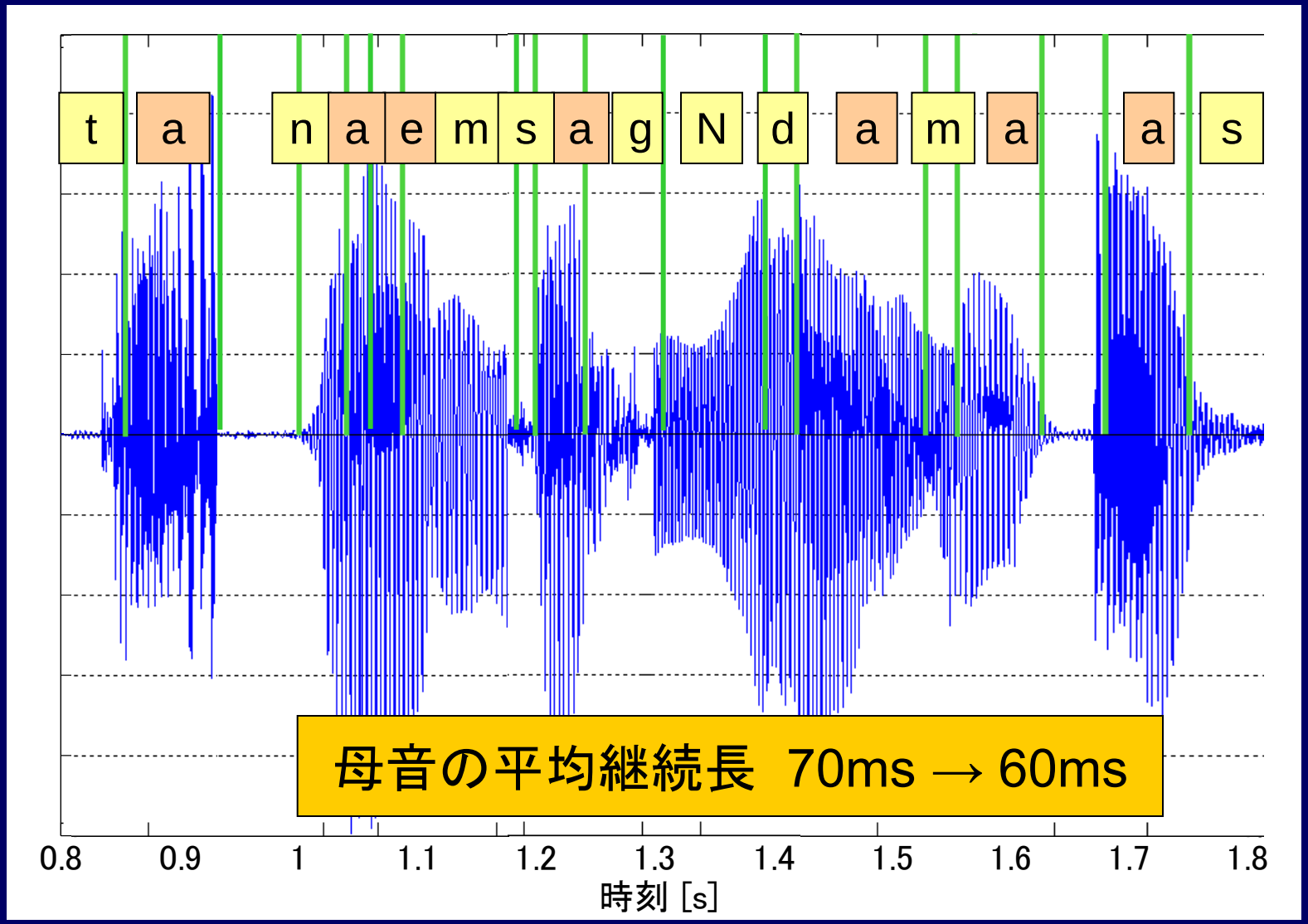
- 歌声の伸ばす発声に着目
- 声の大きさの変動が大きければ歌声
- 女性音声の方が朗読音声と歌声の音高差が大きく識別しやすい
- 音声信号内のF0の変動が大きければ歌声

Filteringした音声に対する感想

- 発声速度，リズムの有無に着目
- 音高が持続する箇所がみられれば歌声
- イントネーションの違いに着目

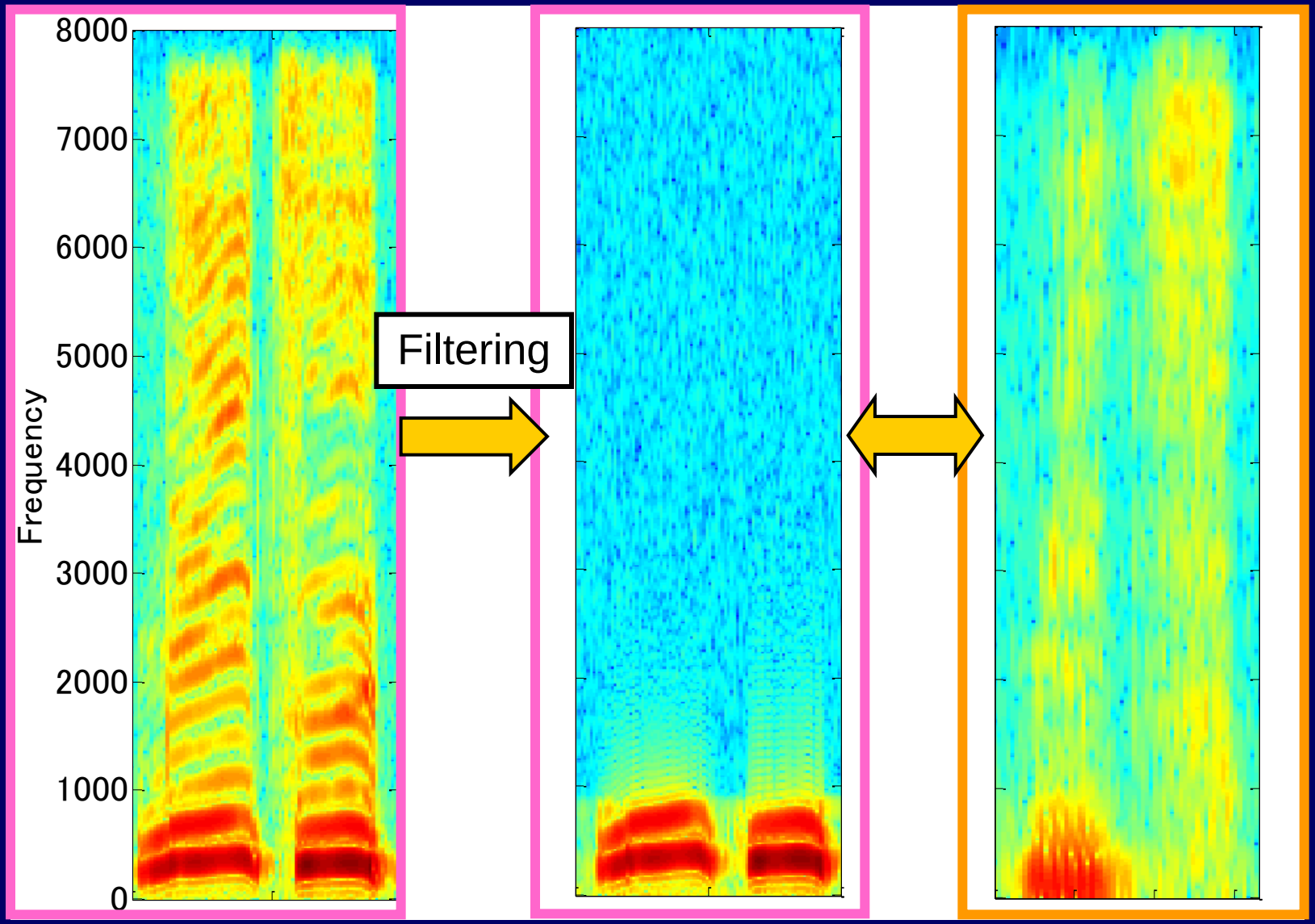
朗読音声の場合

- 音素(母音)の継続長の変化

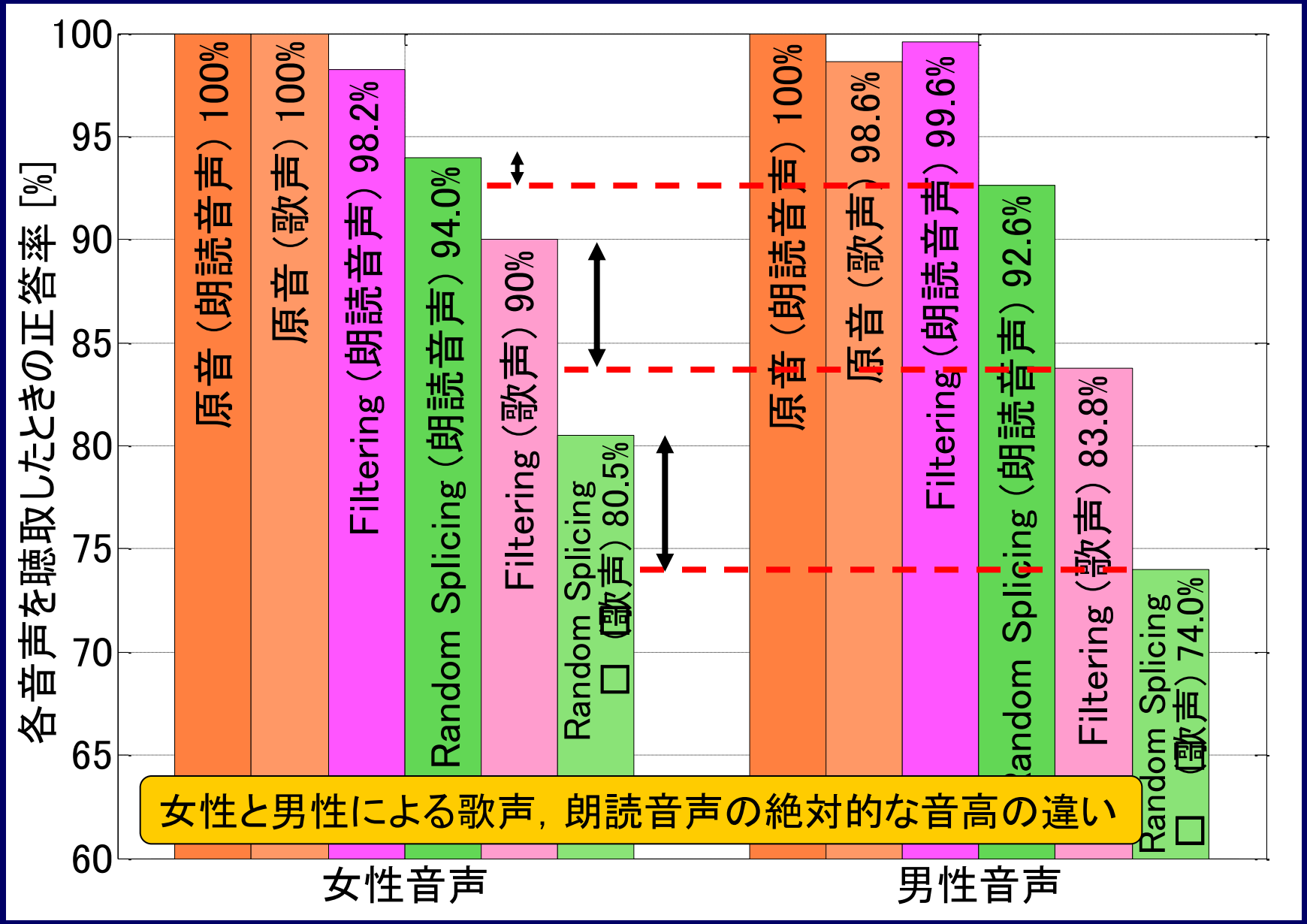


Filtering手法による不正答の考察

歌声　加工音声　朗読音声



性別ごとの音声からみた聴取実験結果



本研究の目的

識別方法

- 言語情報の利用
  音声認識により発声内容から音声を識別
- 非言語情報の利用
  イントネーション，テンポ，音色などから音声を識別

歌の歌い方，話し方というような
発声のスタイルの違いに着目

歌声とは

- 歌声の典型的な特徴
  - 基本周波数(以下, F0と呼ぶ)と強度が幅広く変化
  - *Singing Formant*
    - オペラ歌手の歌声
    - 喉頭の部分で共鳴を起こし, 深い響きを作り出す歌唱法
    - 必ずしも素人の歌声に観測できるとは限らない

人間はたとえ素人の歌声であったとしても
少しの聴取により話し声との識別が可能

    - 発声の長さの違い
    - テンポの違い
    - 音高の変化の違い

従来研究

- 音楽と音声のカテゴリの識別手法
  - 周波数領域の特徴量
    Spectral Centroid, MFCC, Harmonic Coefficient
  - 時間領域の特徴量
    ゼロ交差回数
  - 周波数・時間の両者に着目した特徴量
    Spectral Flux, 4-Hz Modulation Energy

➡ 混合音の音響特徴量の検討

  - 楽器の混合音や伴奏付きの歌声

    歌声そのものの特徴は，まだ十分に議論されていない

本研究の目的

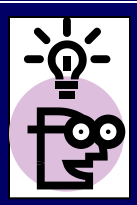
- 歌声と朗読音声の自動識別手法の提案
  - 発声機構による歌声の物理的な声質の明確化
  - 歌い方，話し方という長時間に観測できる発声のスタイルの違い

応用例

- 音声対話システムにおける発話検出
- 音声合成の精度の向上
- 自律型ロボットの聴覚的情景分析
- 歌声，話し声による楽曲検索システム

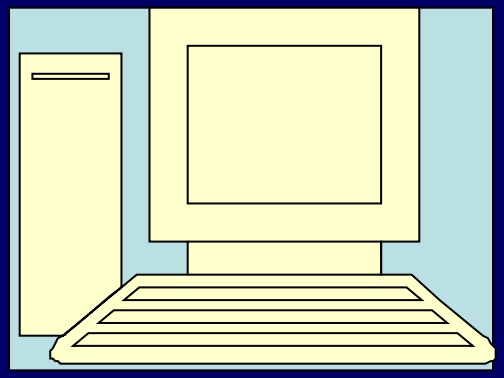
自動音声識別器をもつ楽曲検索システム

話し声



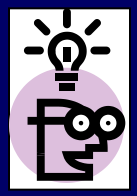
“イブズの「恋のver.2.4」を聞かせてください”

検索システム



「恋のver.2.4」

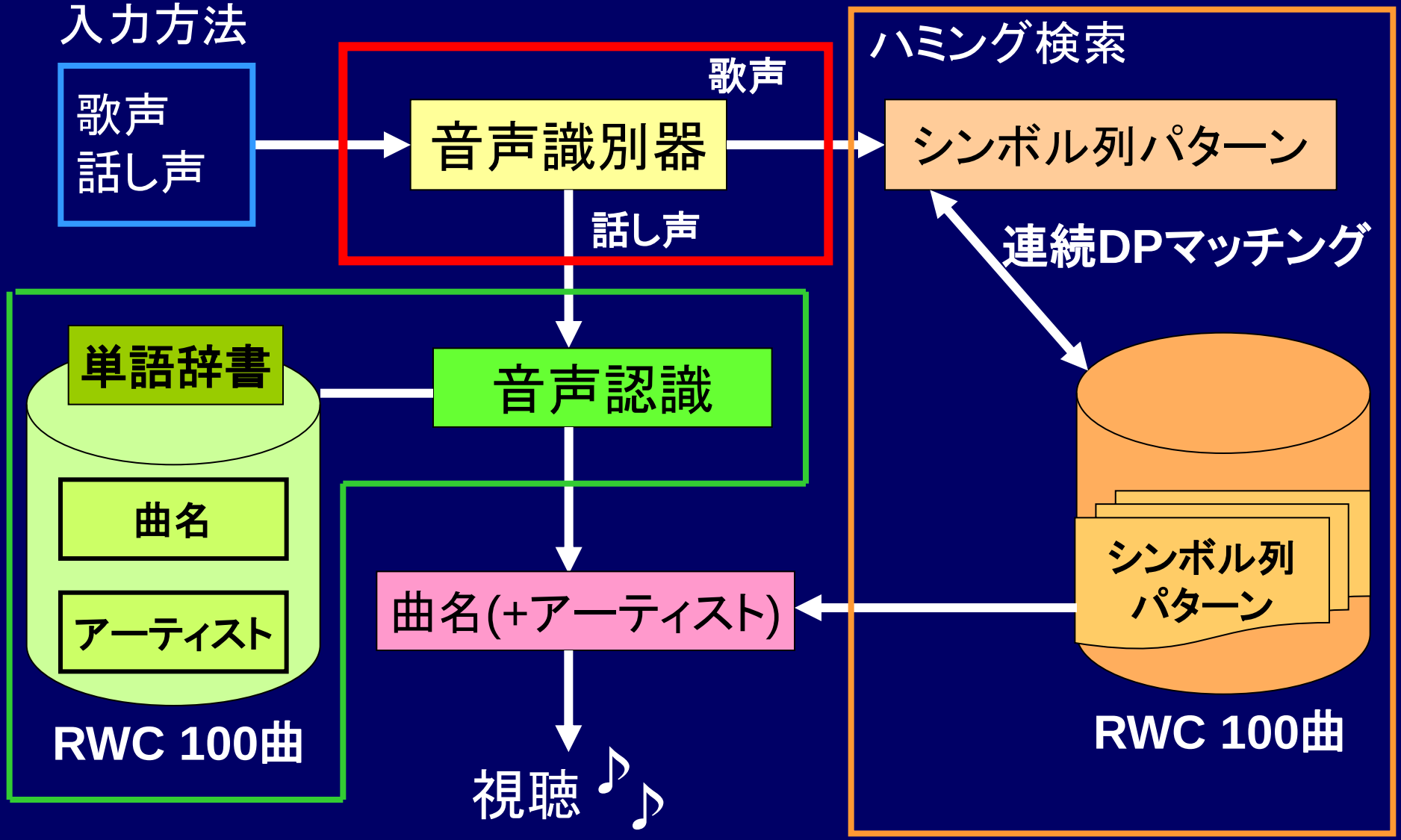
歌声



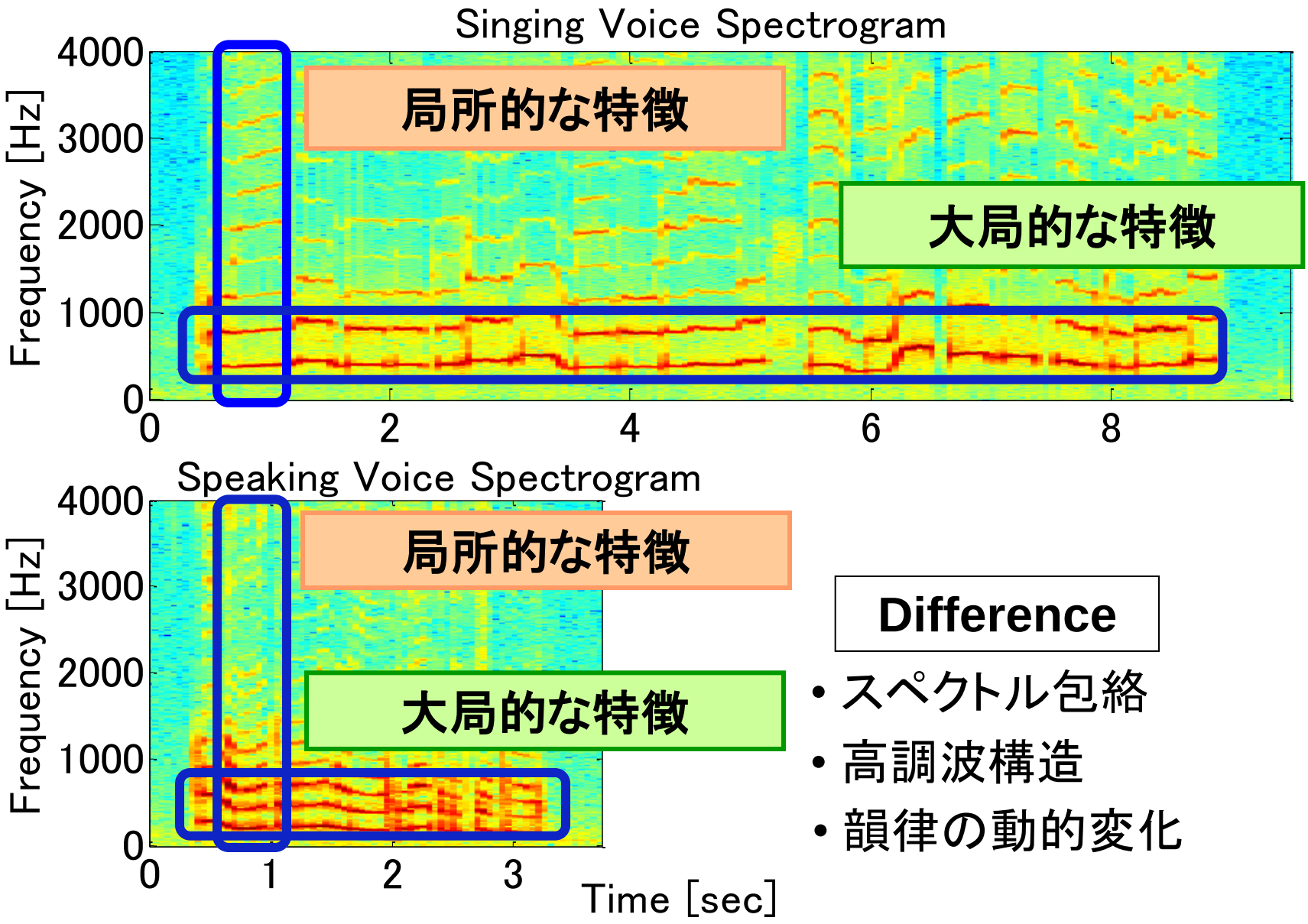
“～線路は続くよ，どこまでも～”

「線路は続くよ」

自動音声識別器をもつ楽曲検索システム



識別特徴量



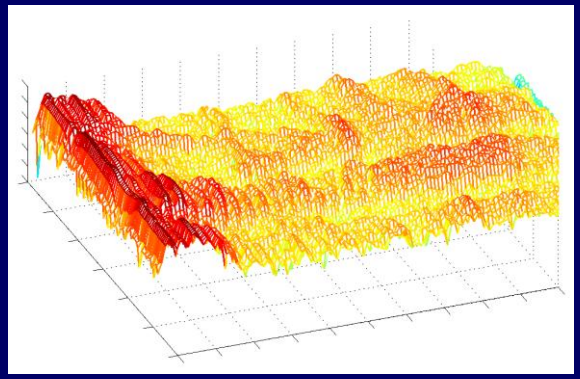
Difference

- スペクトル包絡
- 高調波構造
- 韻律の動的変化

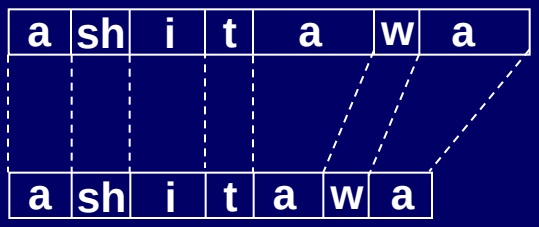
局所的な特徴による尺度

- スペクトル包絡の違い

Mel-Frequency Cepstrum Coefficients (MFCCs)

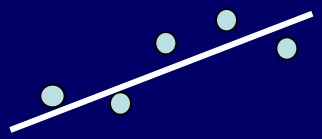


  - 100-msハミング窓を利用
  - 10 msごとに算出
- 母音の長さの違い
  - 歌声 : 伸ばす発声
  - 朗読音声 : 音素が次々と変化



ΔMFCCs (MFCC derivatives)

  - 5点の回帰係数

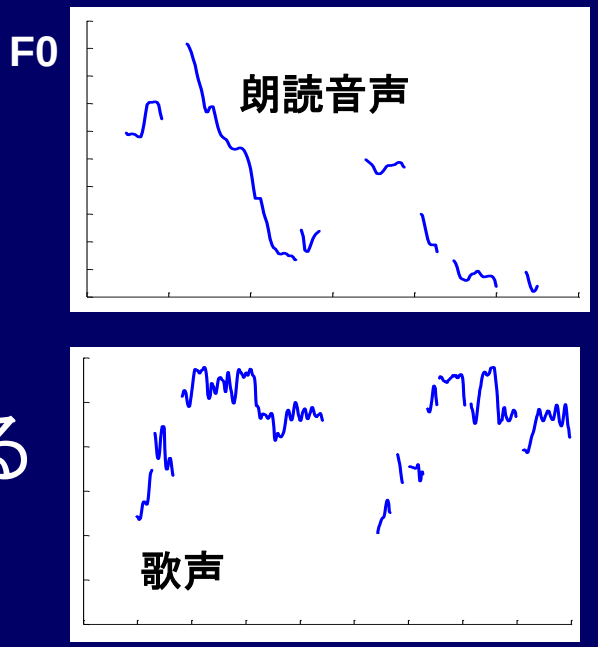


大局的な特徴による尺度

- F0 推定
  - 優勢休止検出のためのF0推定手法(後藤ら)
  - メディアンフィルタによる平滑化
- 韻律の変化の違い

ΔF0 (five-point regression)

  - 朗読音声のF0は下降
  - 歌声は曲のメロディの制約を受ける

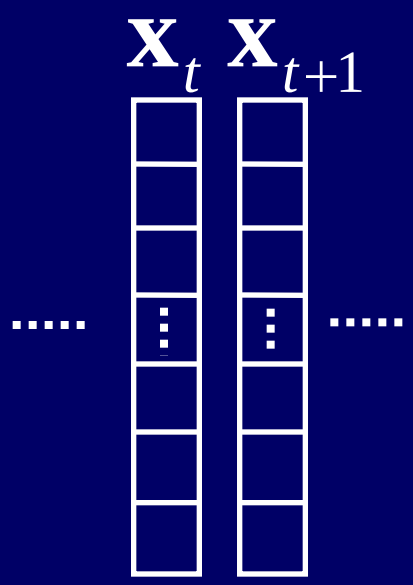


歌声，朗読音声の識別方法

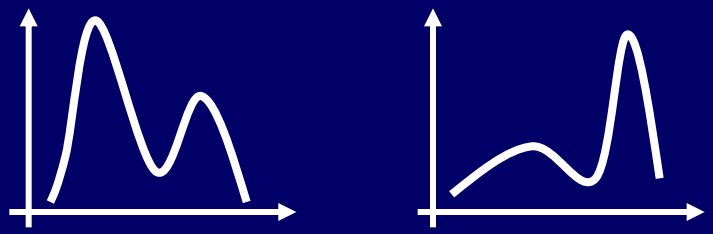
- 16混合ガウス分布(GMM)による識別

入力ベクトル系列 $\mathbf{X}$
(MFCC, ΔMFCC, ΔF0)

$\mathbf{x}_t$ $\mathbf{x}_{t+1}$



識別器



歌声　　朗読音声

$$\hat{d} = \arg\max_{d=\text{歌声,朗読音声}} \sum_{t=1}^{N} \log f(\mathbf{x}_t; \Lambda_d)$$

$\Lambda_d(d = \text{歌声,朗読音声})$ は
MFCC, ΔMFCC, ΔF0ベクトルの
分布に対するGMMのパラメータ

歌声
or
朗読音声