歌声と朗読音声の自動識別のための基本周波数の時間構造のモデル化に関する検討

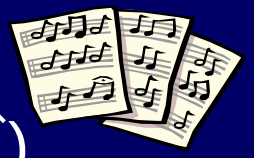
大石 康智[1], 後藤 真孝[2]
伊藤 克亘[1], 武田 一哉[1]
[1]名古屋大学大学院情報科学研究科
[2]産業技術総合研究所

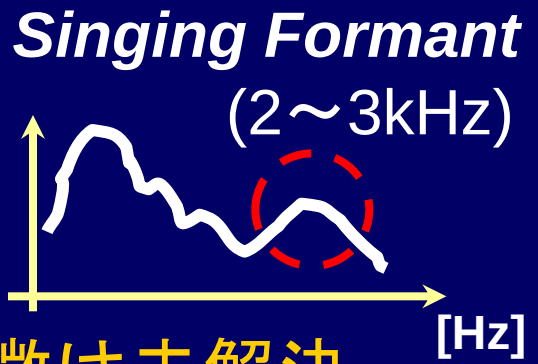
はじめに

音声の発声スタイルの違い　歌声と話し声

歌声は，曲に従って歌い方が変化し，
技能が必要な非常に幅の広い発声スタイル

- 歌声の音響的特徴
  - 音高が幅広く変化 (2オクターブ)
  - ***Singing Formant*** (オペラ歌手)
    - 必ずしも素人の歌声に観測できない





話し声との識別に影響する音響的特徴は未解決

- 話し声と音楽の自動識別
  - 音楽は，混合音(楽器音，伴奏付き歌声)

伴奏なしの歌声と話し声の識別はされていない

本研究の目的

歌声と話し声の音響的特徴分析
自動識別尺度を提案

人間の識別能力を調査するための聴取実験
- 識別に必要な音声信号長，音響的特徴の調査

聴取実験の知見に基づいた自動識別実験

AISTハミングデータベース

- RWC音楽データベースから抜粋した様々なジャンルの計25曲
  - **Aメロ** (楽曲中の歌唱の出だしの部分)
  - **サビ** (一番代表的な盛り上がる主題の部分)

歌唱する(伴奏なし): 歌声

歌詞を読み上げる: 朗読音声

識別能力の調査を目的とした聴取実験

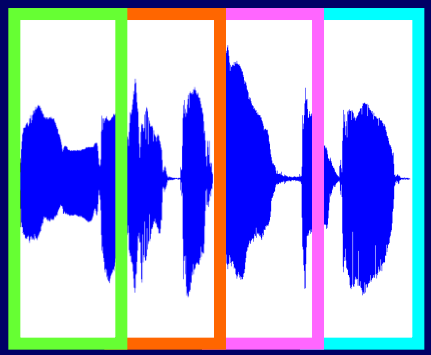
- 識別に必要な音声信号長の調査

  ➡ 1sの音声信号: 100%識別可 (10名の被験者)

- 識別に影響する音響的特徴の調査

  1sの音声信号の時間構造を変形

  - Random Splicing



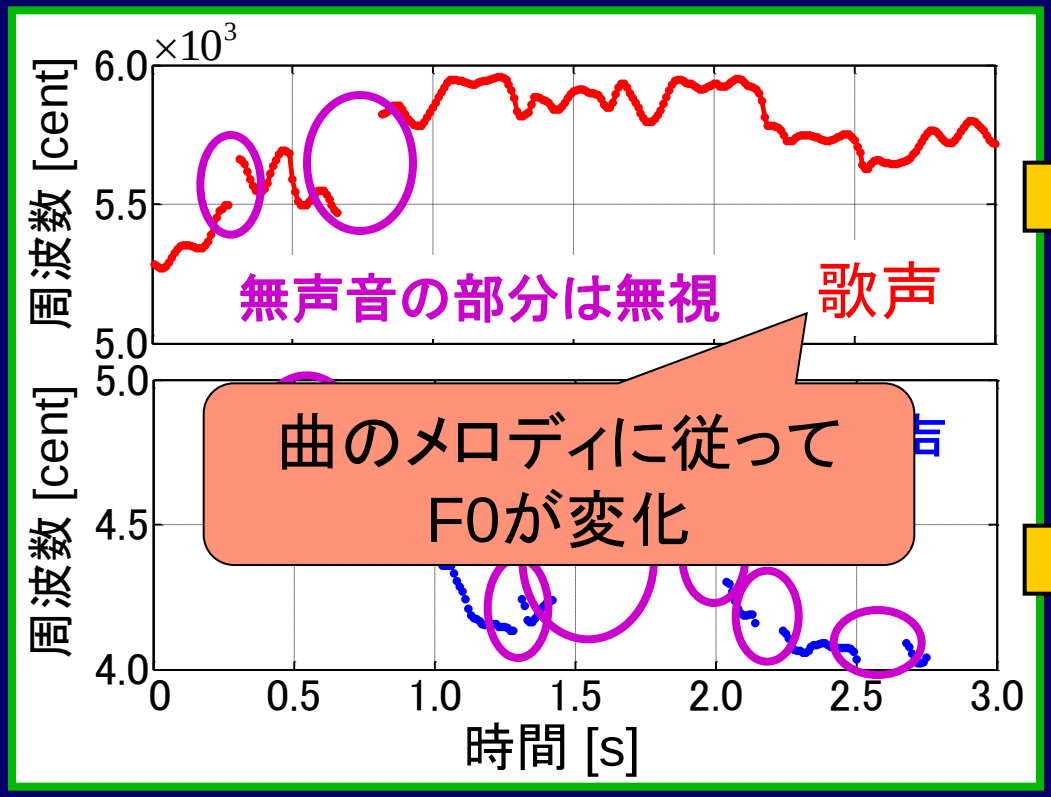
ランダムに
並べ替え

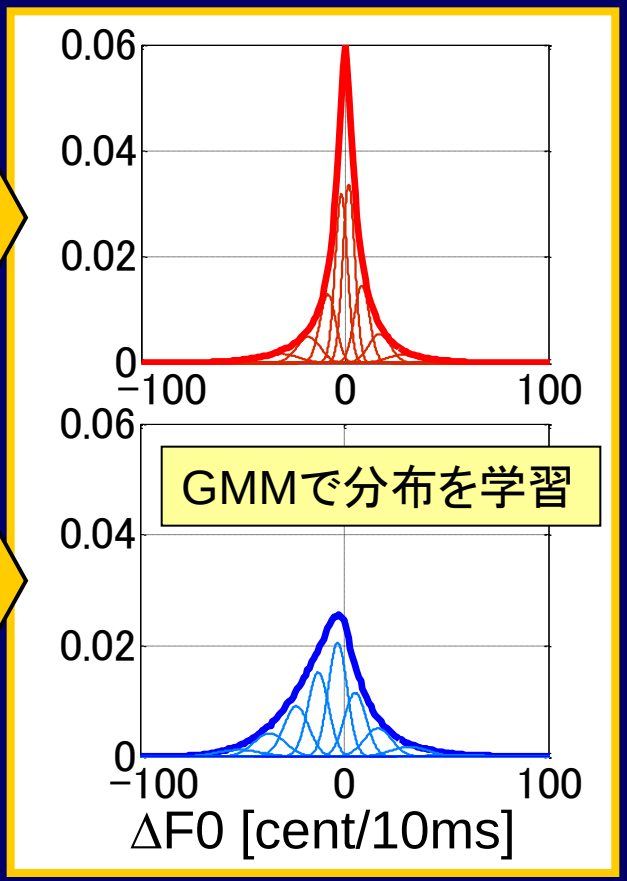
  - 断片長を短くするにつれて歌声の識別率が大きく低下
  - 音素長の変化 (歌声の平均母音長:146.7ms → 73.3ms)
    (朗読音声の平均母音長:70.0ms → 60.0ms)

従来の自動識別尺度

- 音声信号の時間的に変化する特徴が重要

基本周波数の時間変化(ΔF0)



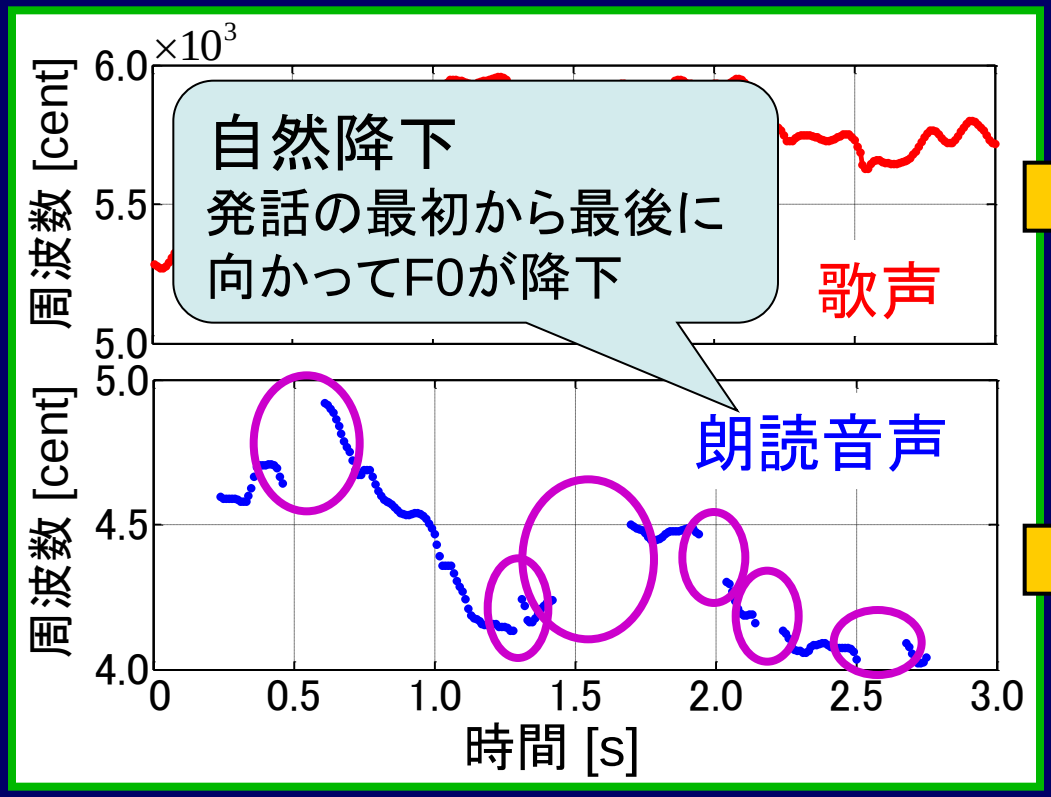


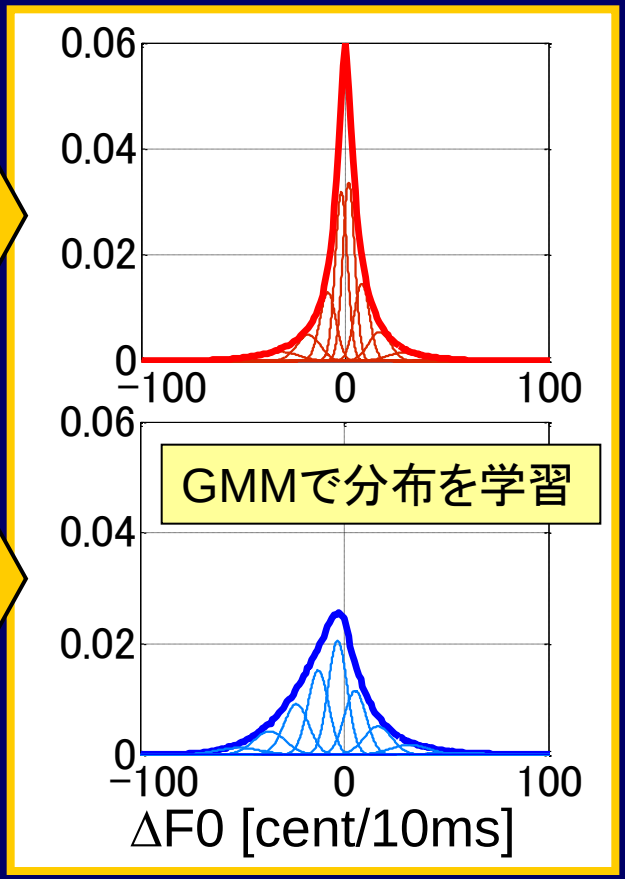
F0の推定には有声休止検出のための
F0推定手法(後藤ら，2000)を利用

従来の自動識別尺度

- 音声信号の時間的に変化する特徴が重要

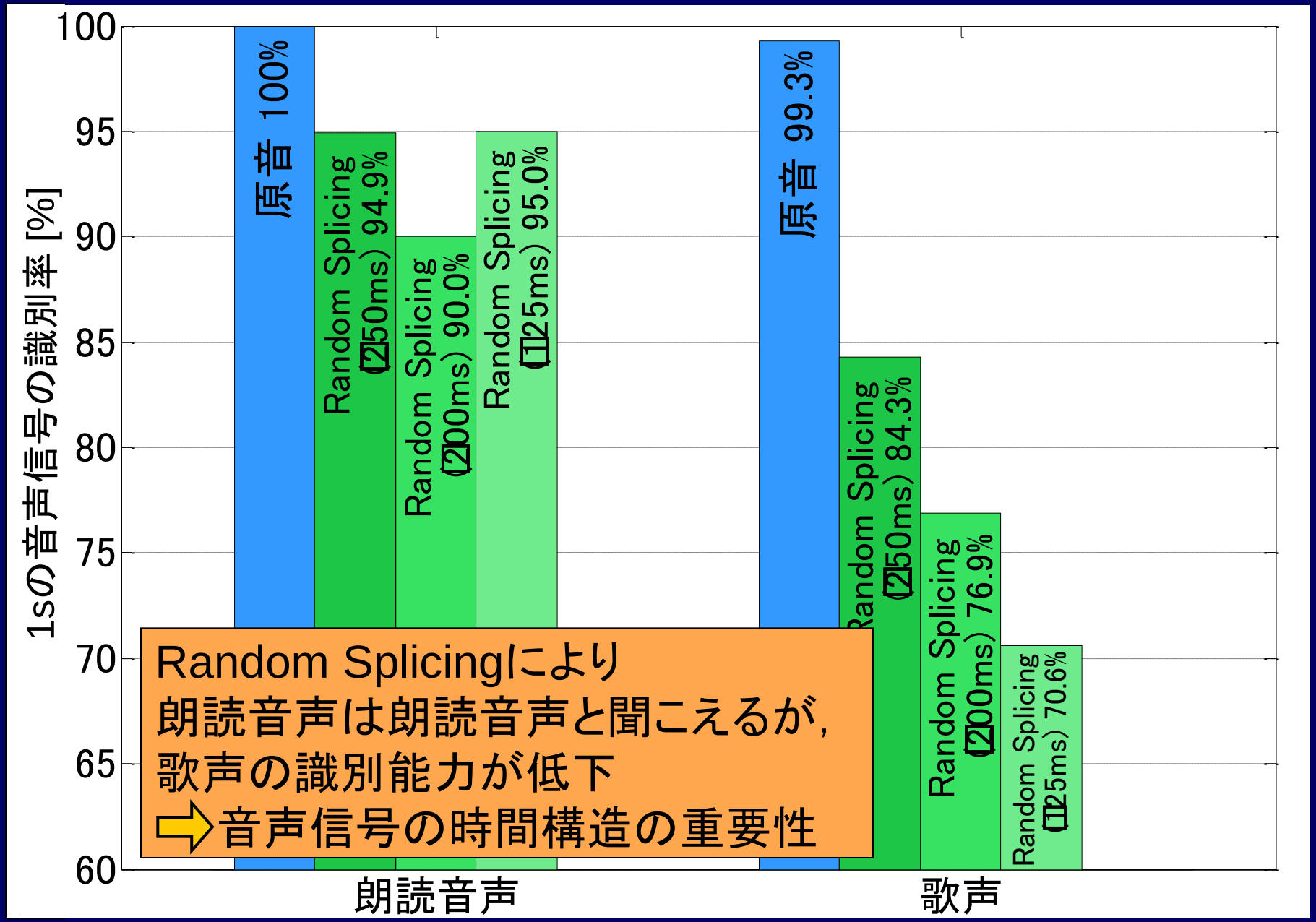
基本周波数の時間変化(ΔF0)



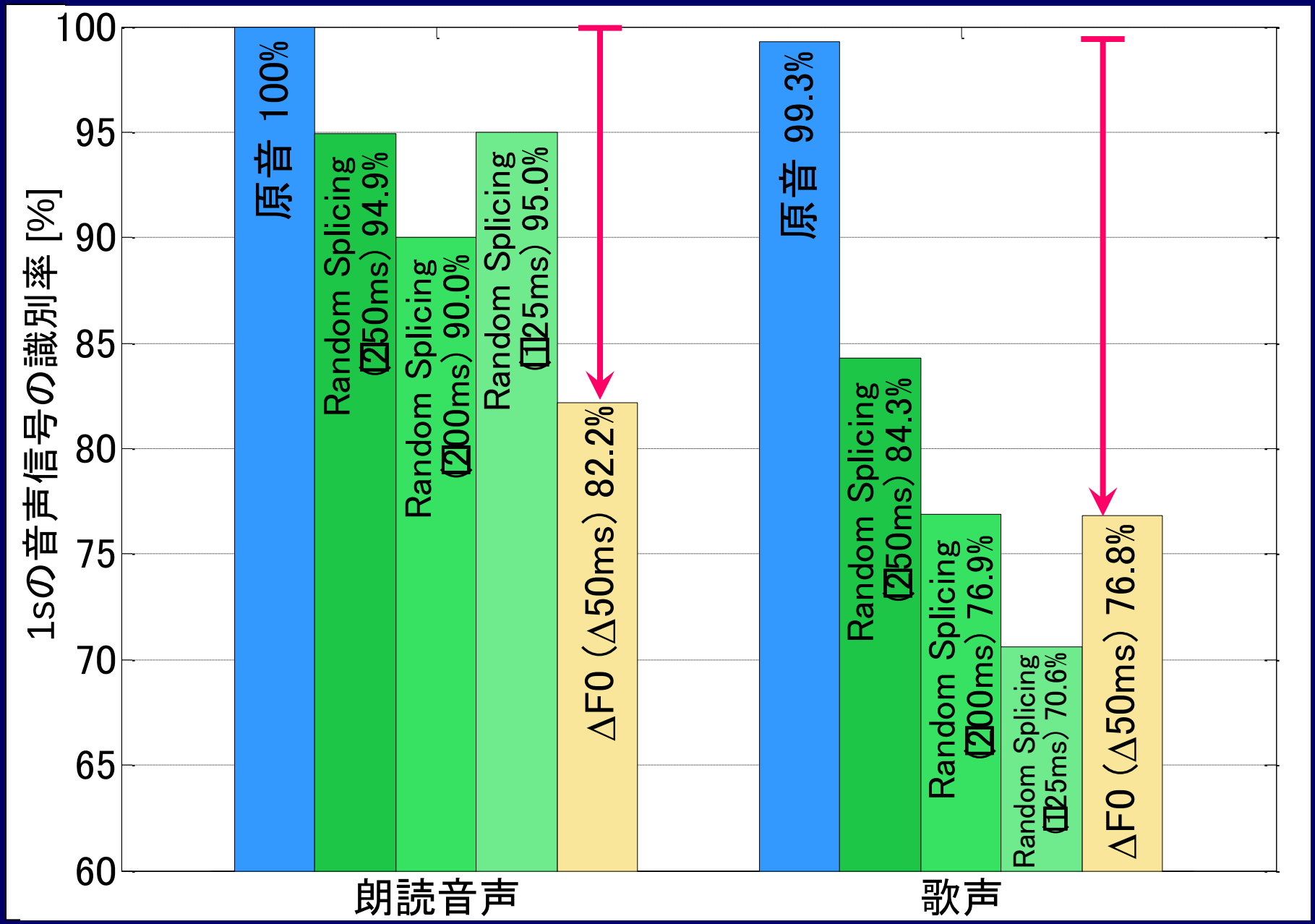


F0の推定には有声休止検出のための
F0推定手法(後藤ら，2000)を利用

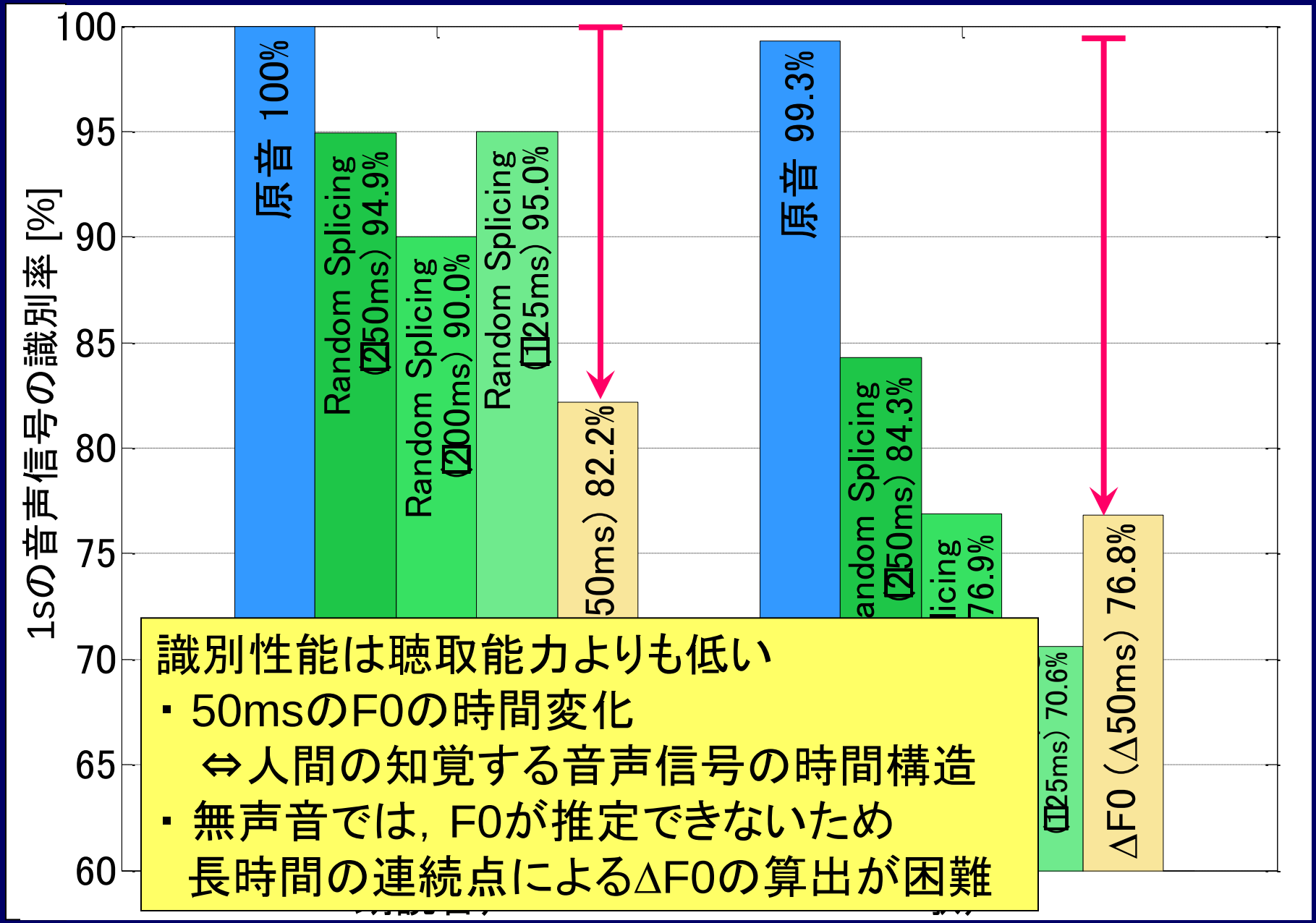
聴取実験と従来の自動識別結果



聴取実験と従来の自動識別結果



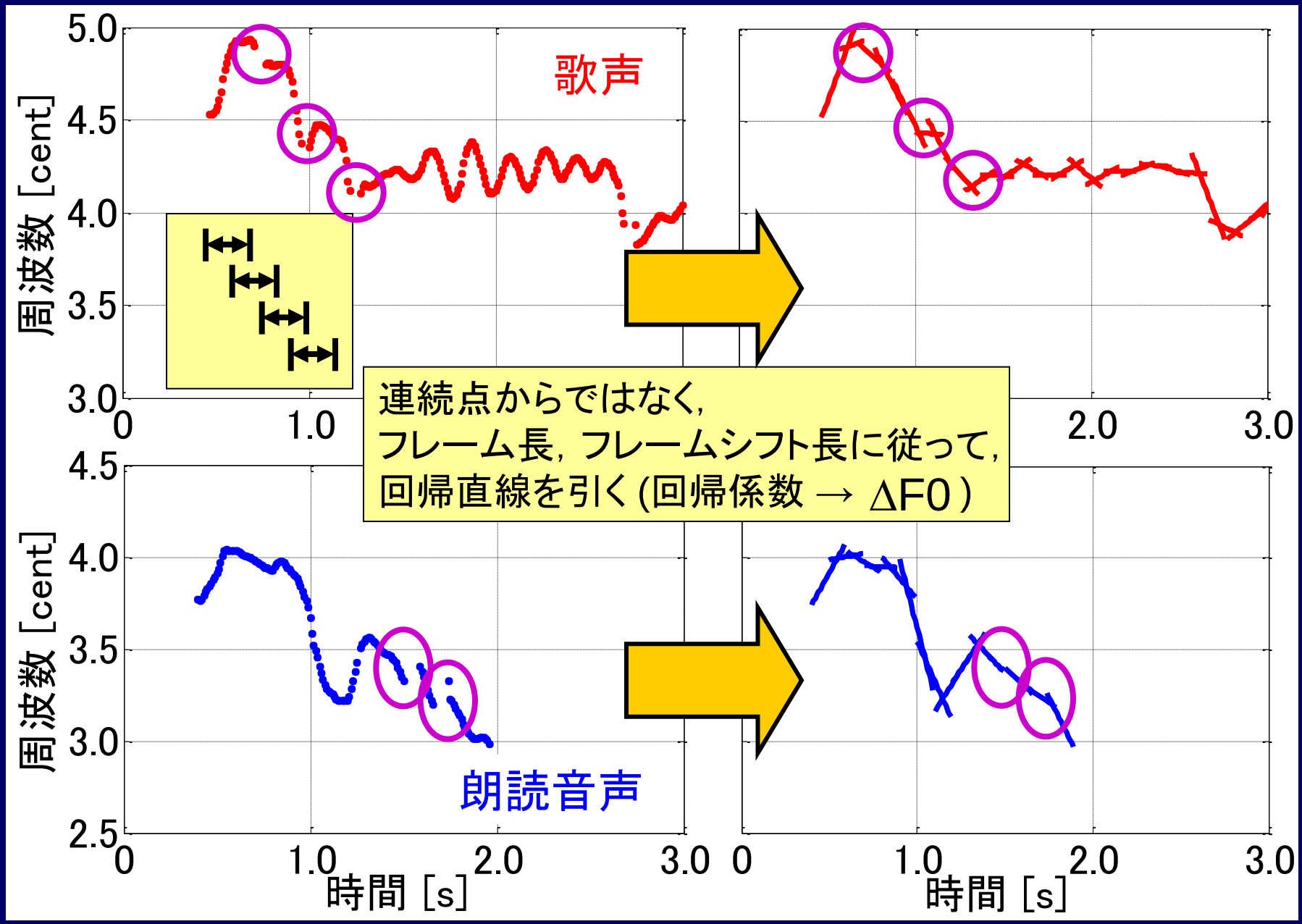
聴取実験と従来の自動識別結果



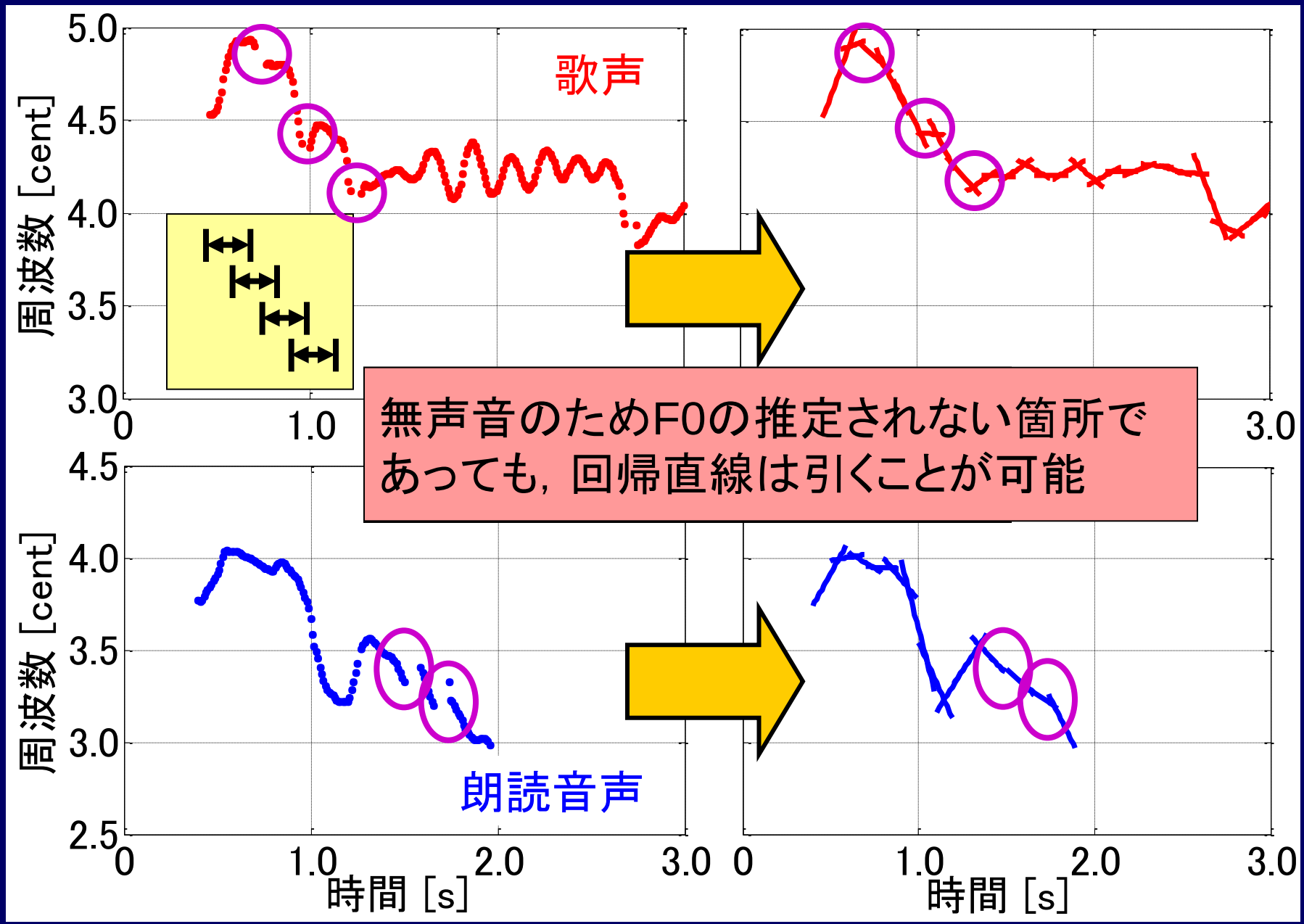
識別性能は聴取能力よりも低い

- 50msのF0の時間変化
  ⇔人間の知覚する音声信号の時間構造
- 無声音では，F0が推定できないため
  長時間の連続点によるΔF0の算出が困難

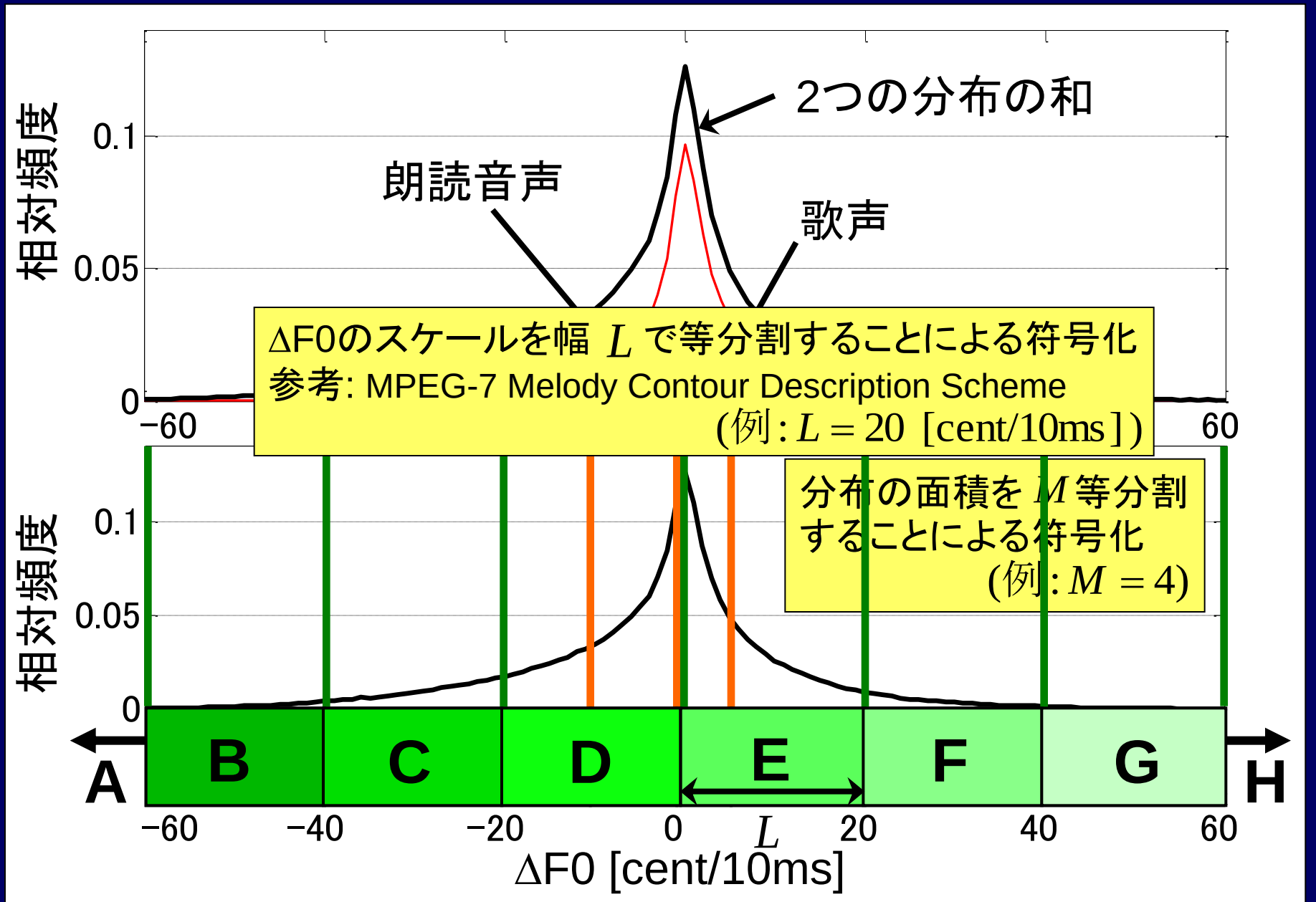
ΔF0の算出方法の変更



ΔF0の算出方法の変更



2つの基準による ΔF0の符号化



符号語列のN-gramモデルによる学習と評価

符号化流れ:

符号語列　A　C　D　R　B　C　C

学習: 歌声と朗読音声のN-gramモデルを学習 (N = 1, 2, 3)

1-gram: 符号語の生起確率　$P_d(\mathrm{A}), P_d(\mathrm{B}), \ldots P_d(\mathrm{R})$

2-gram:

$P_d(\mathrm{A}\,|\,\mathrm{A}), P_d(\mathrm{B}\,|\,\mathrm{A}), \ldots$

3-gram:

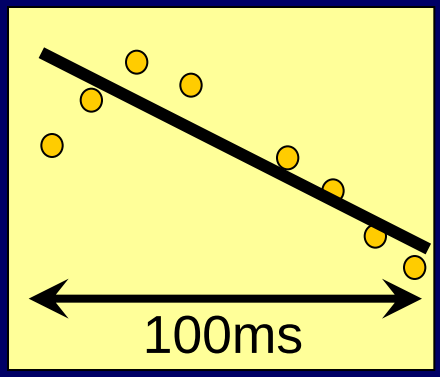
$P_d(\mathrm{A}\,|\,\mathrm{AA}),\ P_d(\mathrm{B}\,|\,\mathrm{AA}),\ \ldots$

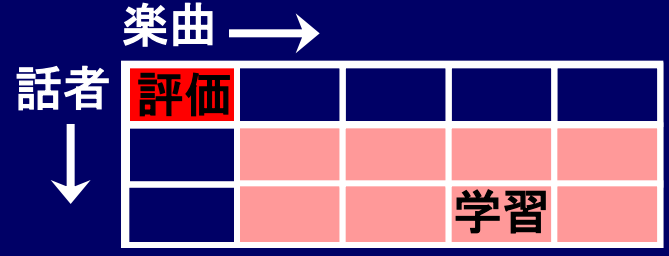
評価:　$\hat{d} = \arg\max P_d$(評価する符号語列)

d=歌声，朗読音声

評価実験

- 使用データ
  - AISTハミングデータベース
    (25曲のAメロとサビ，男性 37名，女性 38名)
- ΔF0の算出パラメータ
  - フレーム長 100ms，フレームシフト長 50ms
- 評価方法
  - オープンデータで評価するために，話者3グループ，
    楽曲5グループに分け，15回のクロスバリデーションを行う

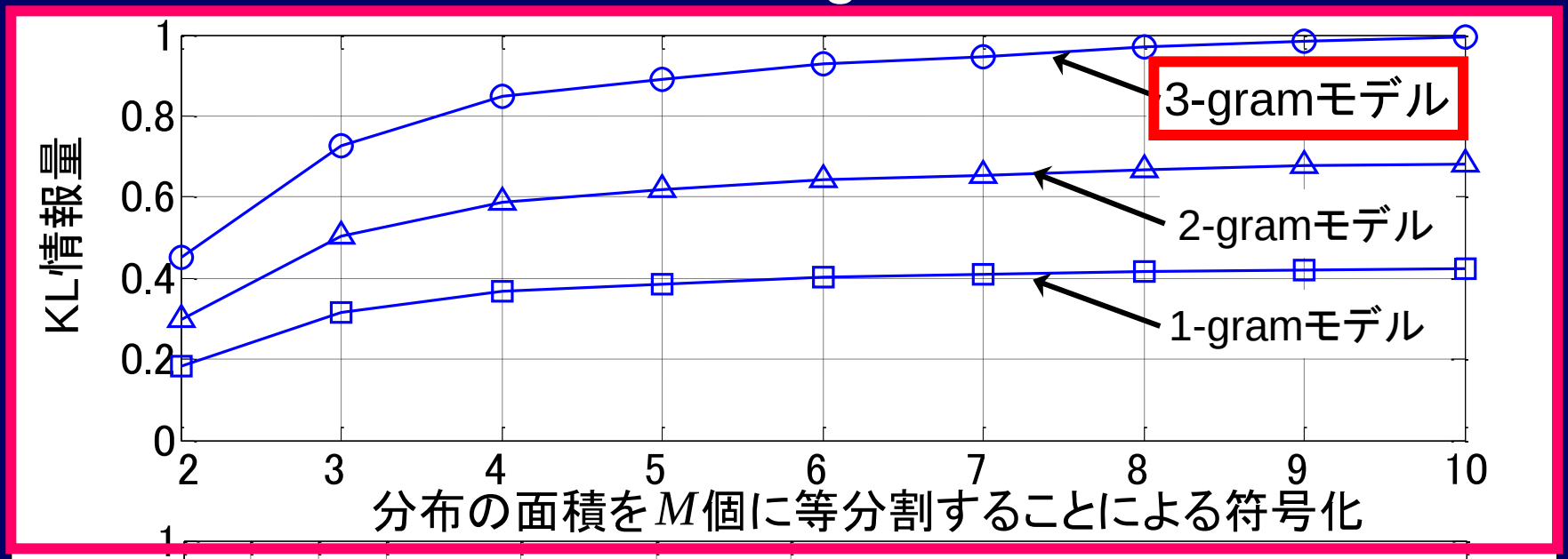


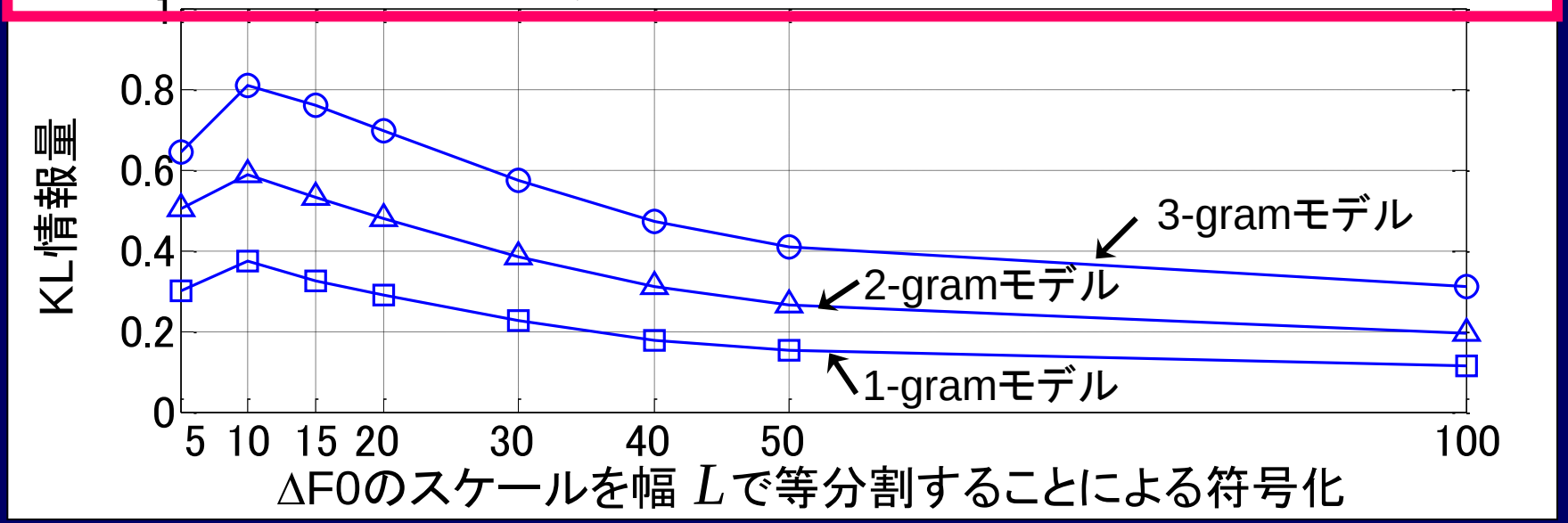


  - ΔF0の符号化方法
    - 分布の面積を M 等分割することによる符号化
    - ΔF0のスケールを幅 L で等分割することによる符号化

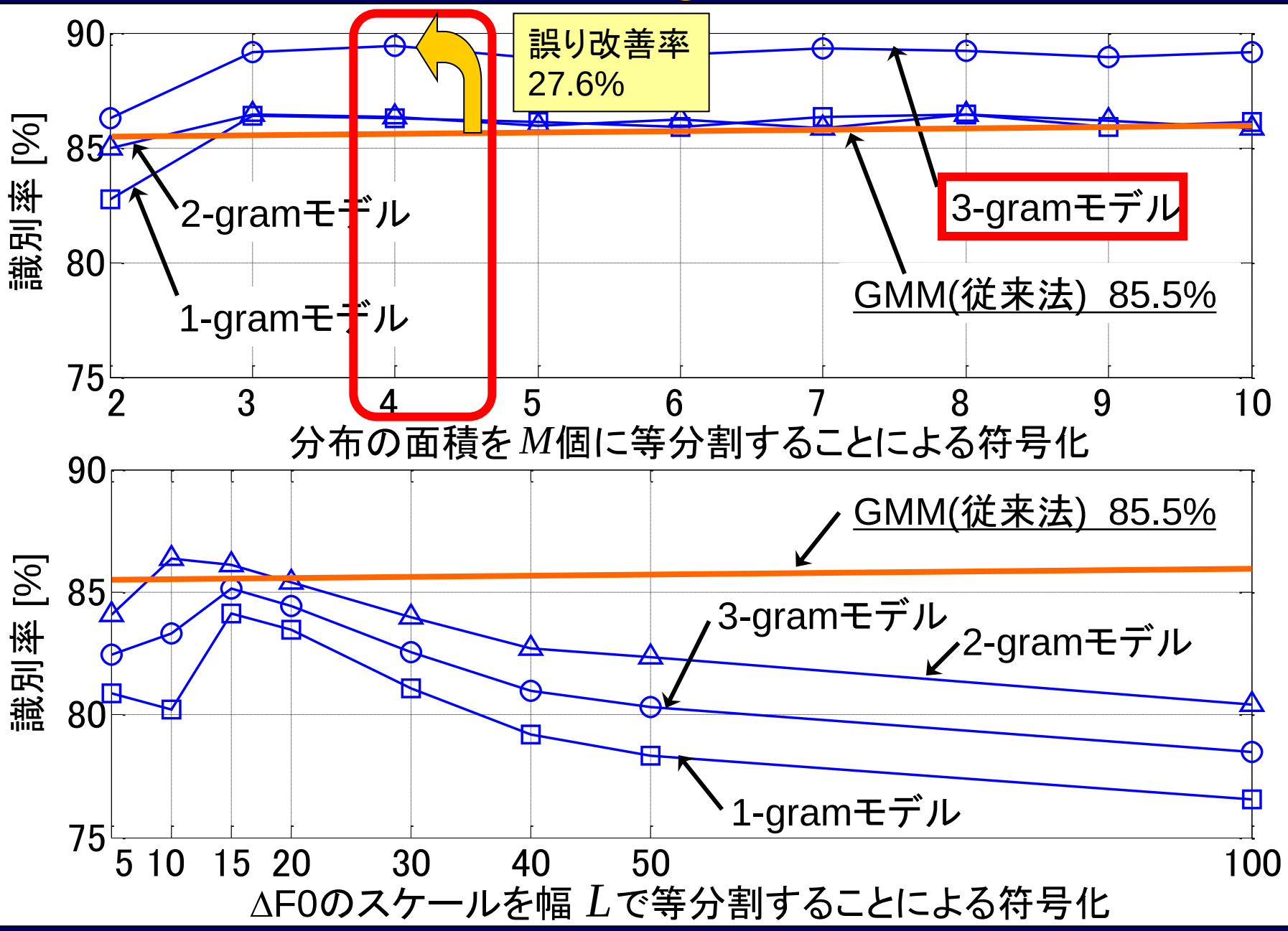
N-gramモデルのKL情報量による評価

- KL情報量が大きい　⇔　N-gramモデルの差が大きい





2sの歌声と朗読音声のN-gramモデルによる識別



まとめと今後の展開

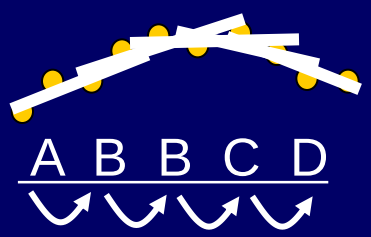
- 歌声と朗読音声のF0の時間構造のモデル化の検討

  人間の音声の識別には時間的に変化する特徴が重要

  歌声：曲のメロディに従って，F0が変化する

  朗読音声：発話の最初から最後に向かって徐々にF0が降下

  従来よりも長時間のΔF0を算出し，その系列を
  文字列に符号化し，N-gramモデルで学習



2sの音声信号に対して，識別率は89.5%

GMMによる従来法に対して，27.6%の誤り改善

- F0以外の時間的に変化する特徴の利用
  (音声信号のパワー，発声速度)