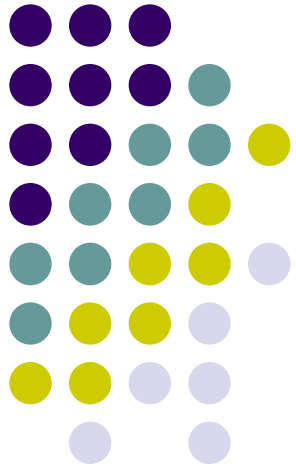
局所的・大局的な特徴を利用した歌声と朗読音声の識別

大石 康智[1], 後藤 真孝[2]
伊藤 克亘[1], 武田一哉[1]

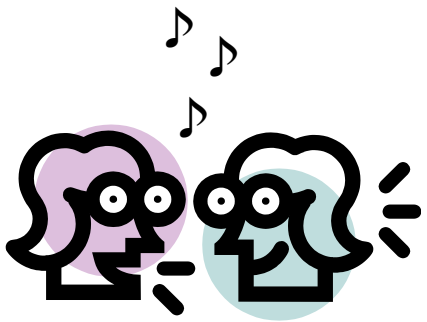
[1]名古屋大学大学院情報科学研究科
[2]産業技術総合研究所



はじめに

- 歌声と話し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



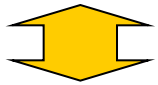
識別方法

- 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

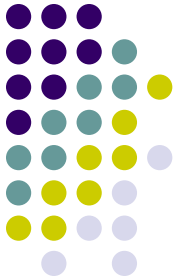
    歌声そのものの特徴は，まだ十分に議論されていない

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

- 歌声と朗読音声の人間の識別能力の調査
- 識別尺度の提案
- 評価実験
- 実験結果の考察
- まとめ

歌声と朗読音声の人間の識別能力の調査



- 識別に必要な音声信号長の調査
- 歌声研究用音楽データベース
  「AISTハミングデータベース」
  - 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い, 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

聴取実験

- 音声サンプル
  - 女性25名, 男性25名

    歌声(計2,500サンプル), 朗読音声(計2,500サンプル)
  - 発声開始から10段階の異なる長さで切り出したもの
    計50,000サンプルを用意
  - 切り出す長さごとにランダムに選択

切り出し時間長	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

歌声と朗読音声の人間の識別能力の調査

- 被験者: 10名(男性9名, 女性1名)
- 各被験者に全430サンプルをランダムな順番で1回だけ聴取
- 回答方法:「歌声」,「朗読音声」,「識別不可能」
  - 聴取音声の例

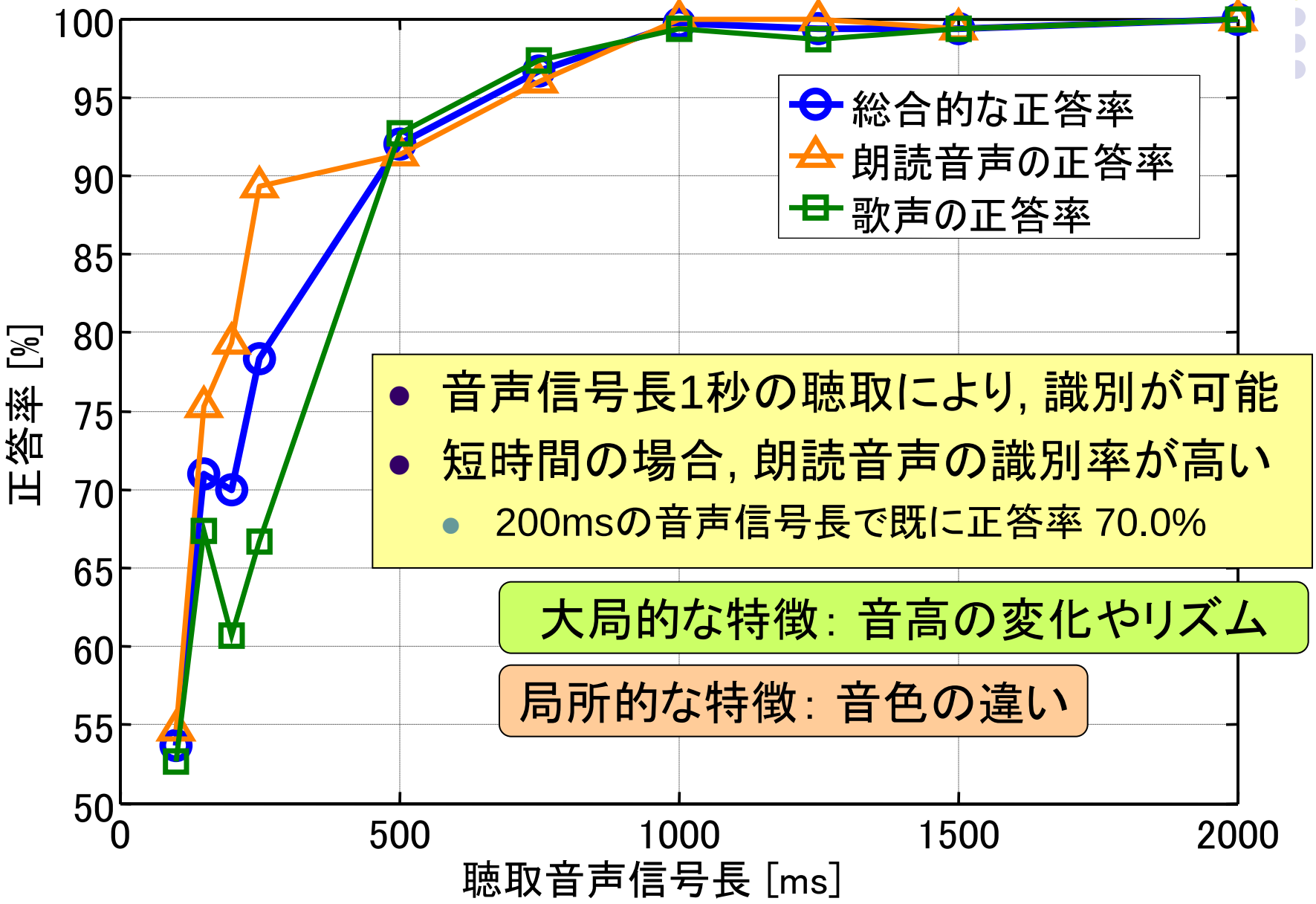
    100msの歌声　　朗読音声

    250msの歌声　　朗読音声

    1000msの歌声　　朗読音声

    2000msの歌声　　朗読音声

人間の識別能力の調査結果

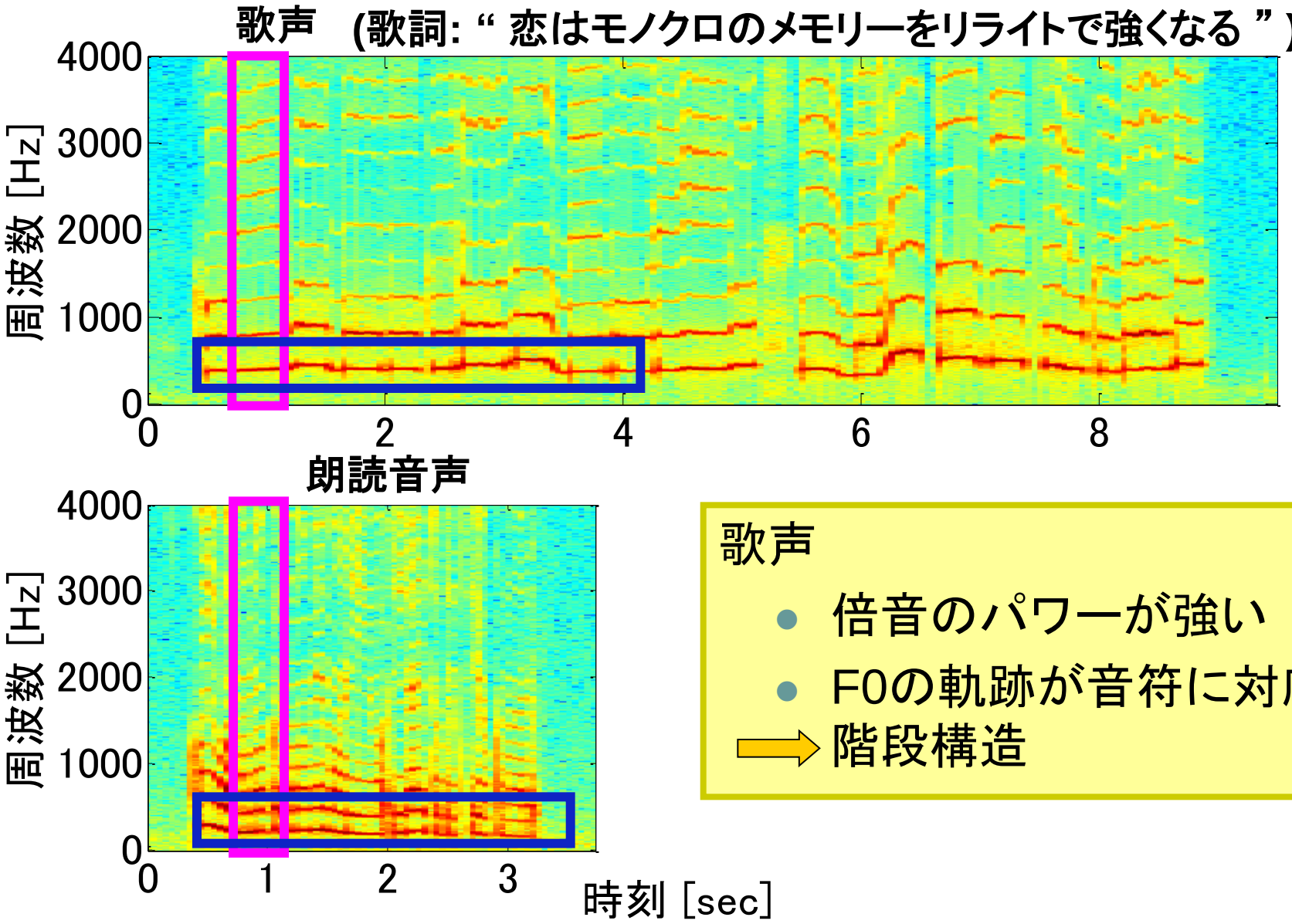


- 音声信号長1秒の聴取により，識別が可能
- 短時間の場合，朗読音声の識別率が高い
  - 200msの音声信号長で既に正答率 70.0%

大局的な特徴：音高の変化やリズム

局所的な特徴：音色の違い

歌声と朗読音声の識別尺度



歌声

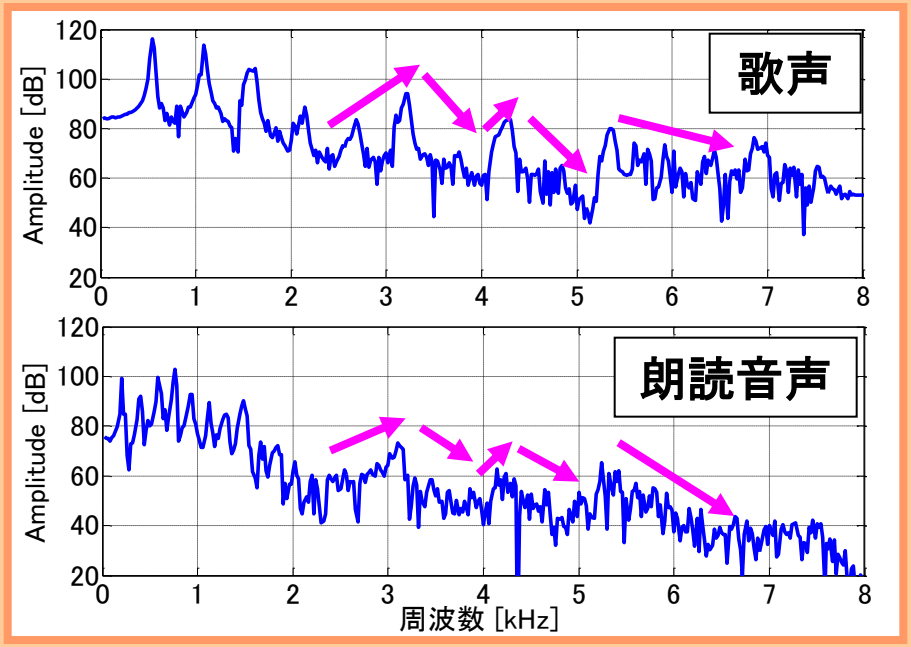
- 倍音のパワーが強い
- F0の軌跡が音符に対応

⇒ 階段構造

歌声と朗読音声の識別尺度

局所的特徴

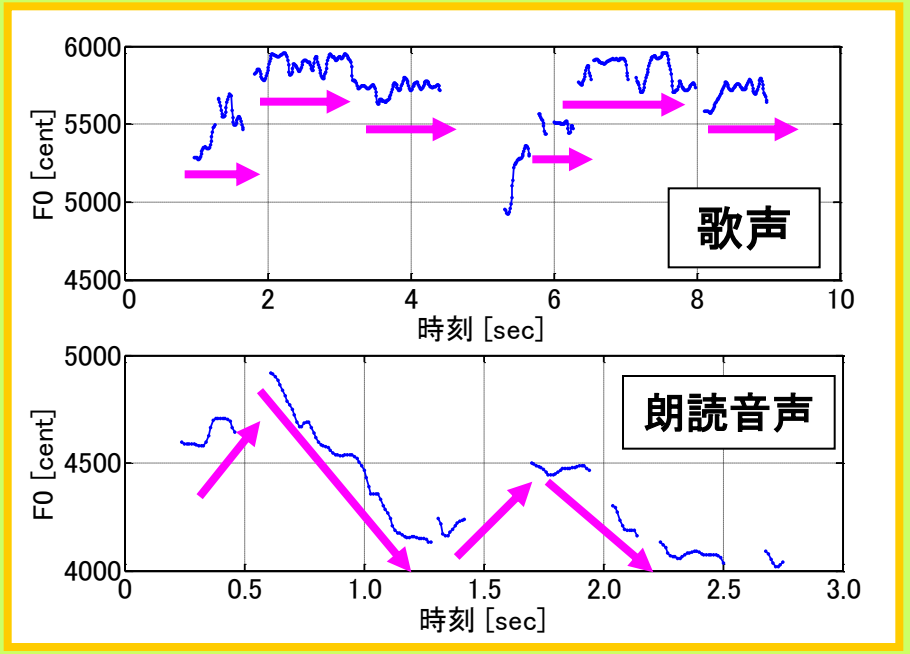
「あ」の発声を25msのフレーム幅で周波数分析



音色，スペクトル包絡の違い

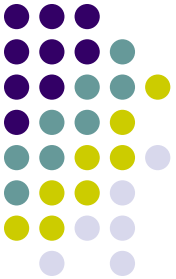
大局的特徴

音声のF0の軌跡



音高の変化や発声長の違い

音声信号の局所的な特徴に基づく尺度



- メル周波数ケプストラム係数(MFCC)
  - スペクトル包絡の違い
- MFCCの時間変化(ΔMFCC)
  - 歌声の伸ばす発声 ⇒ スペクトル包絡の変化: 小

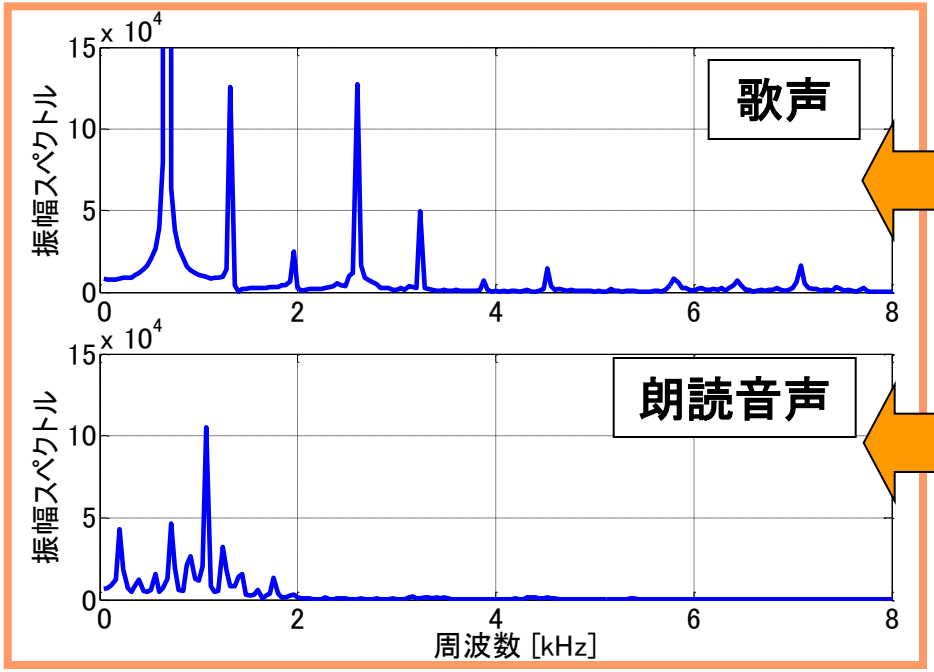
標本化周波数	16kHz
分析窓	ハミング窓
フレームシフト長	10ms
フレーム長	25ms
フィルタバンク数	24
使用帯域	0～8000Hz

⇒ MFCC12次までの係数を利用

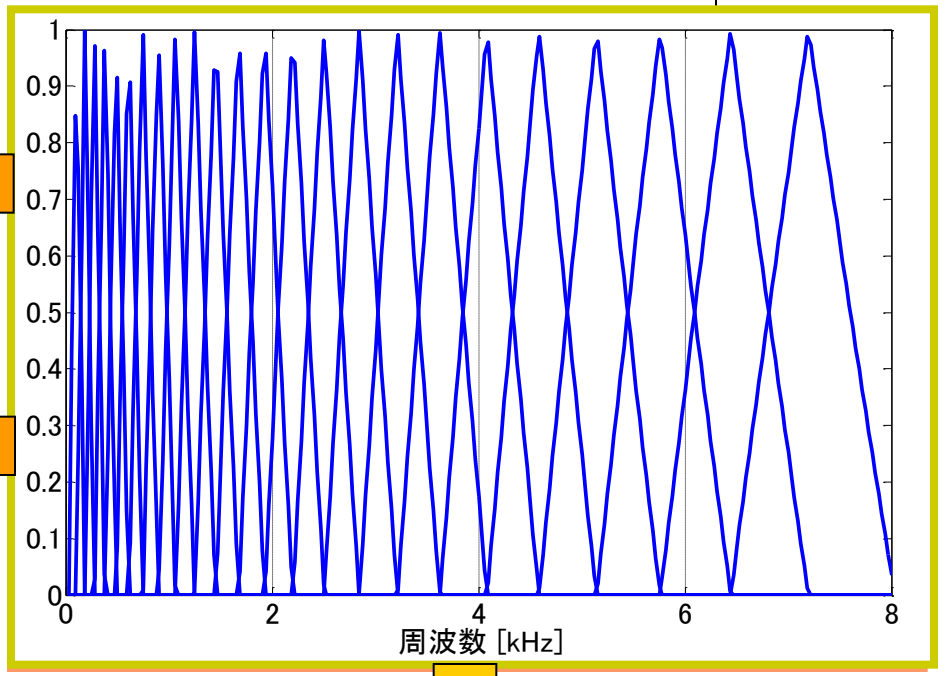
ケプストラム分析

フィルタバンク処理

25msのフレーム幅で周波数分析



24個のフィルタバンクを配置



離散コサイン変換

ΔMFCCの算出 (K=2)

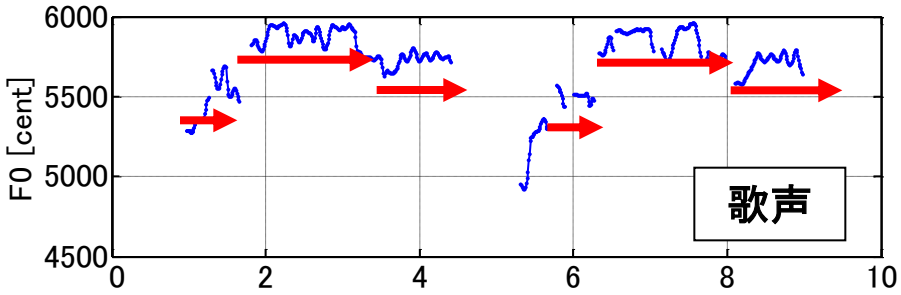
$$\Delta c_n[t] = \frac{\sum_{k=-K}^{K} k \cdot c_n[t+k]}{\sum_{k=-K}^{K} k^2}$$

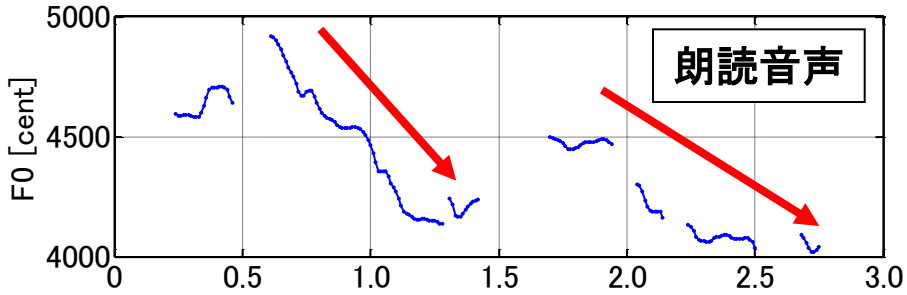
$n = 1, 2, \cdots, 12$ (MFCC係数)

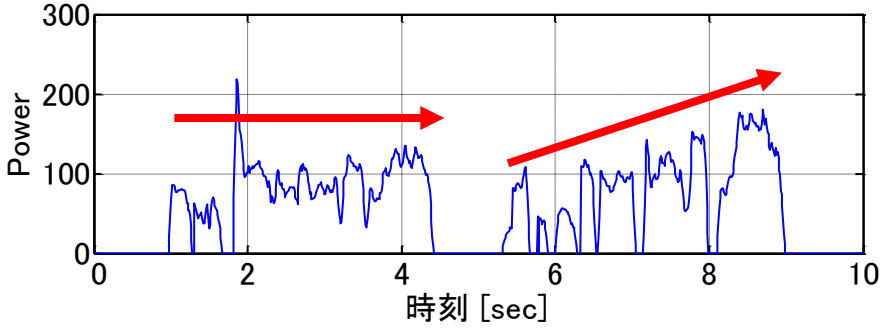
音声信号の大局的な特徴に基づく尺度

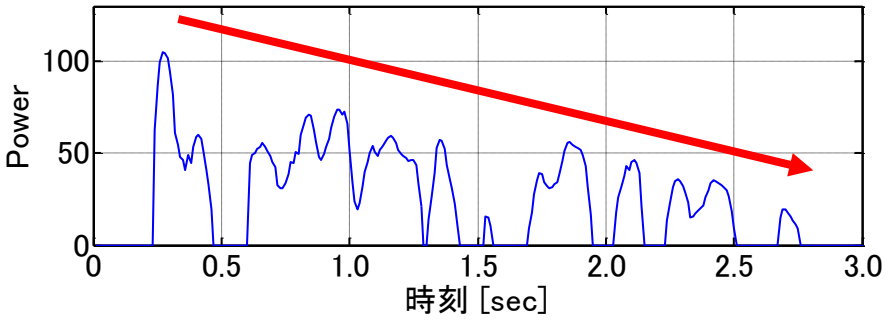
- F0の遷移の違い

歌声は曲のメロディとリズムの制約を受ける
日本語の朗読音声の韻律は下降する

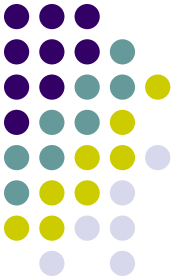








音声信号の大局的な特徴に基づく尺度



- F0の時間変化ΔF0を利用
- F0の抽出
  - 自然発話中の有声休止箇所の検出のために提案されたF0推定手法を利用 (後藤ら, 2004)
  - 10msecで抽出
  - [Hz]で与えられる周波数の単位を[cent]に変換
- ΔF0：各発声区間で5点(50ms)の回帰係数

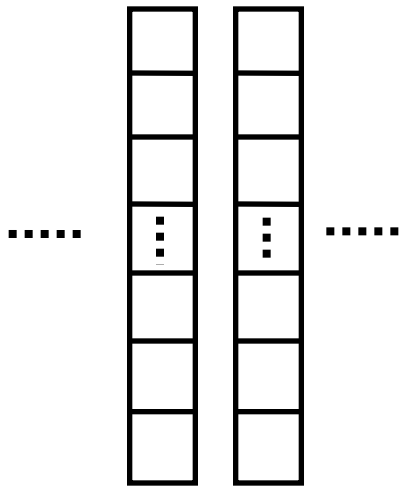
歌声，朗読音声の識別方法

- 16混合ガウス分布(GMM)による識別

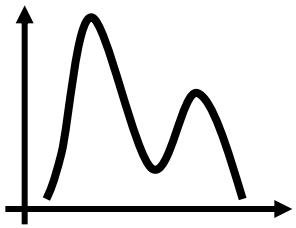
入力ベクトル系列 $\mathbf{X}$

(MFCC, ΔMFCC, ΔF0)

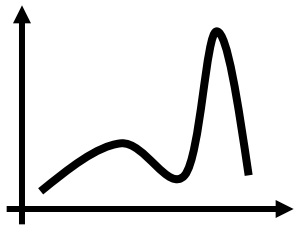
$\mathbf{x}_t$ $\mathbf{x}_{t+1}$



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

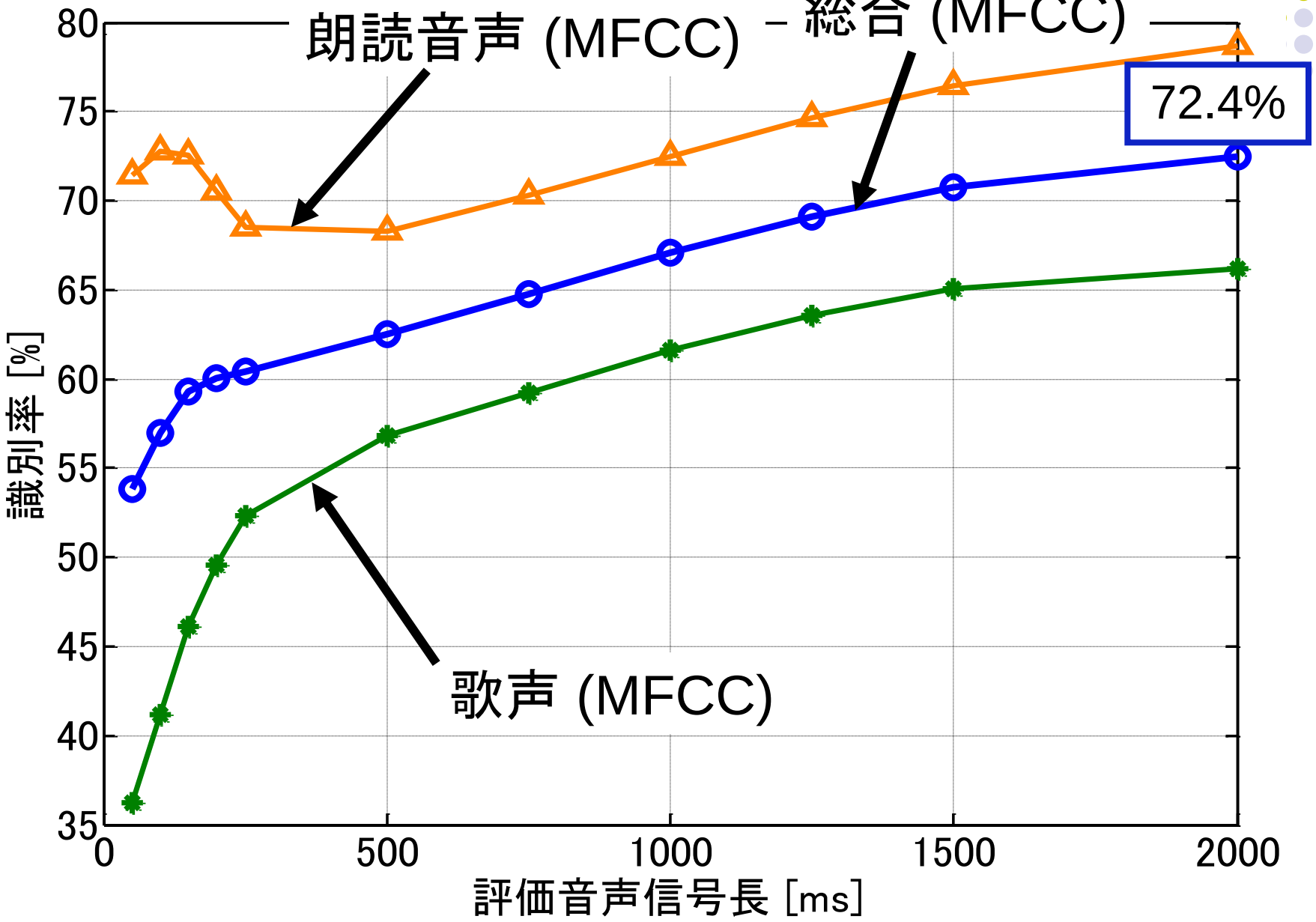
分布に対するGMMのパラメータ

歌声

or

朗読音声

局所的な特徴を利用した識別性能



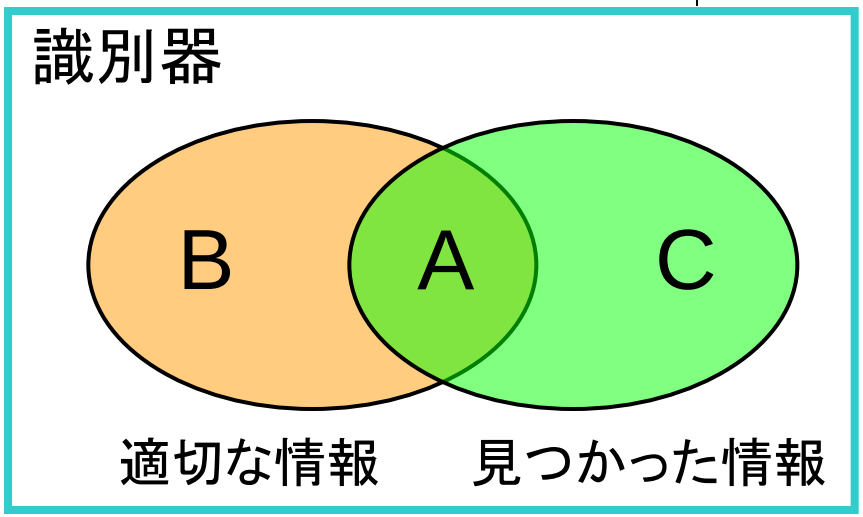
識別性能の評価方法

(例) 歌声の性能の算出

A：歌声→歌声

B：歌声→朗読音声

C：朗読音声→歌声



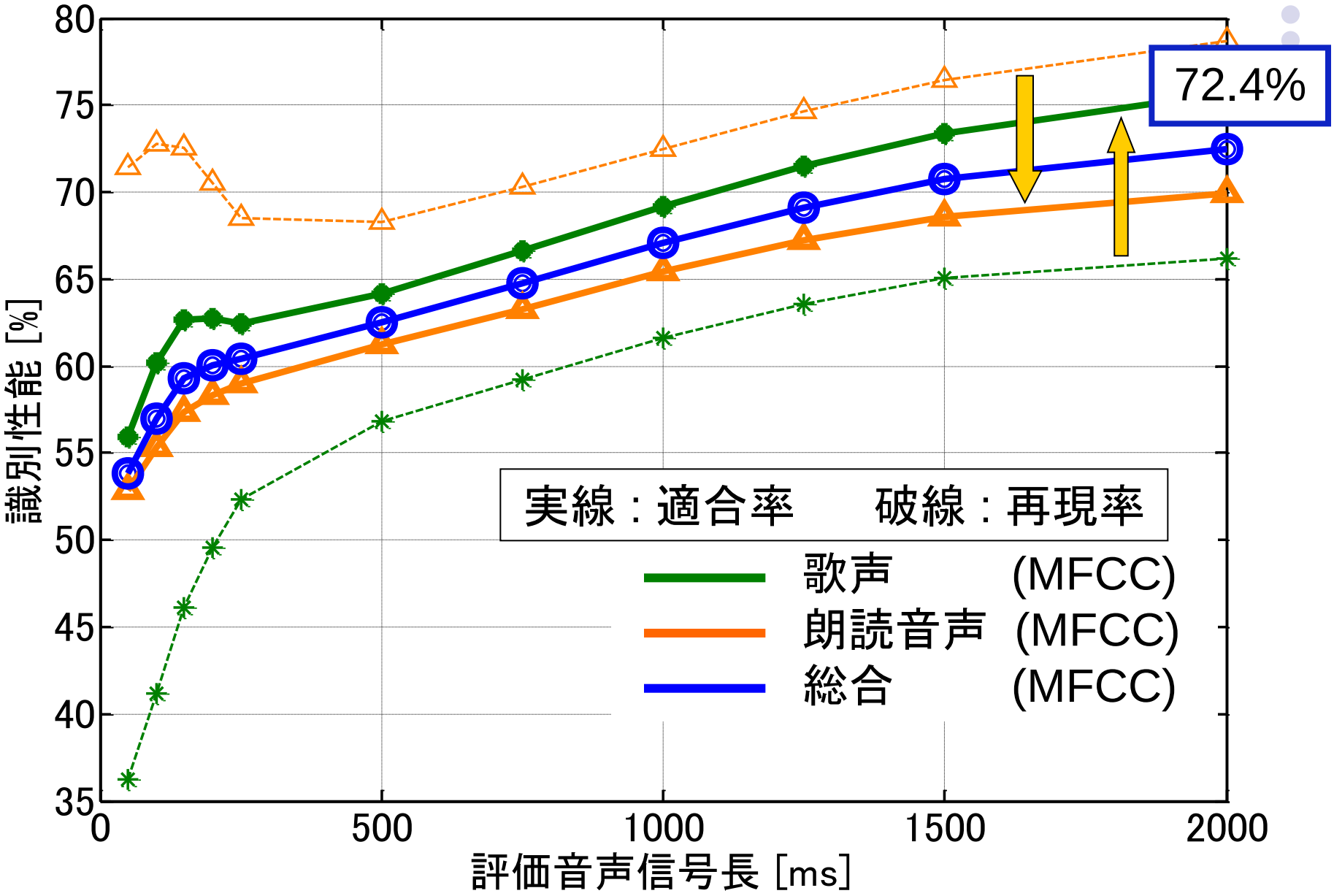
- 再現率(Recall)

  A/(A+B) ⇒ 歌声の識別率

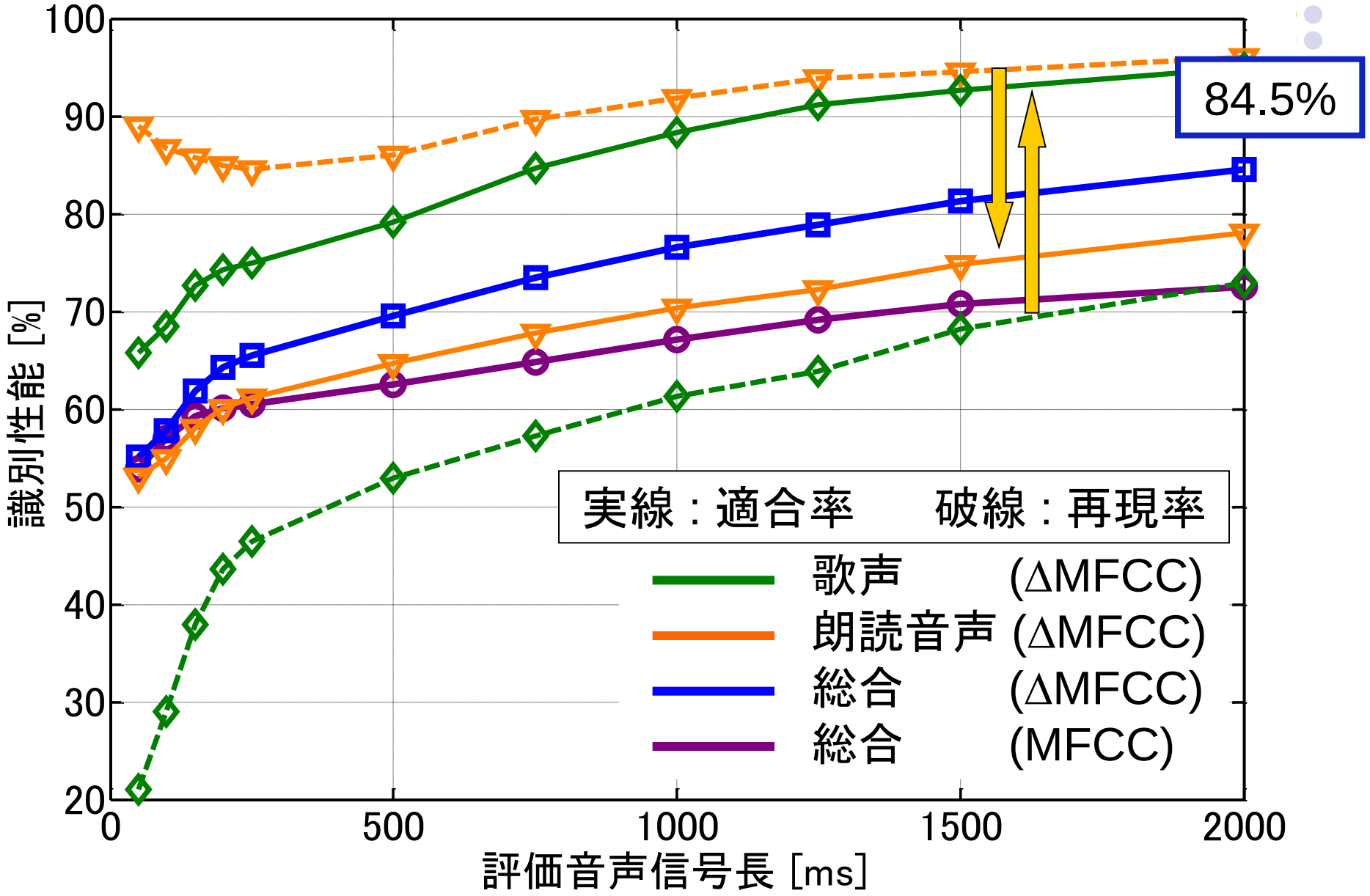
- 適合率(Relevancy)，精度(Precision)

  A/(A+C)

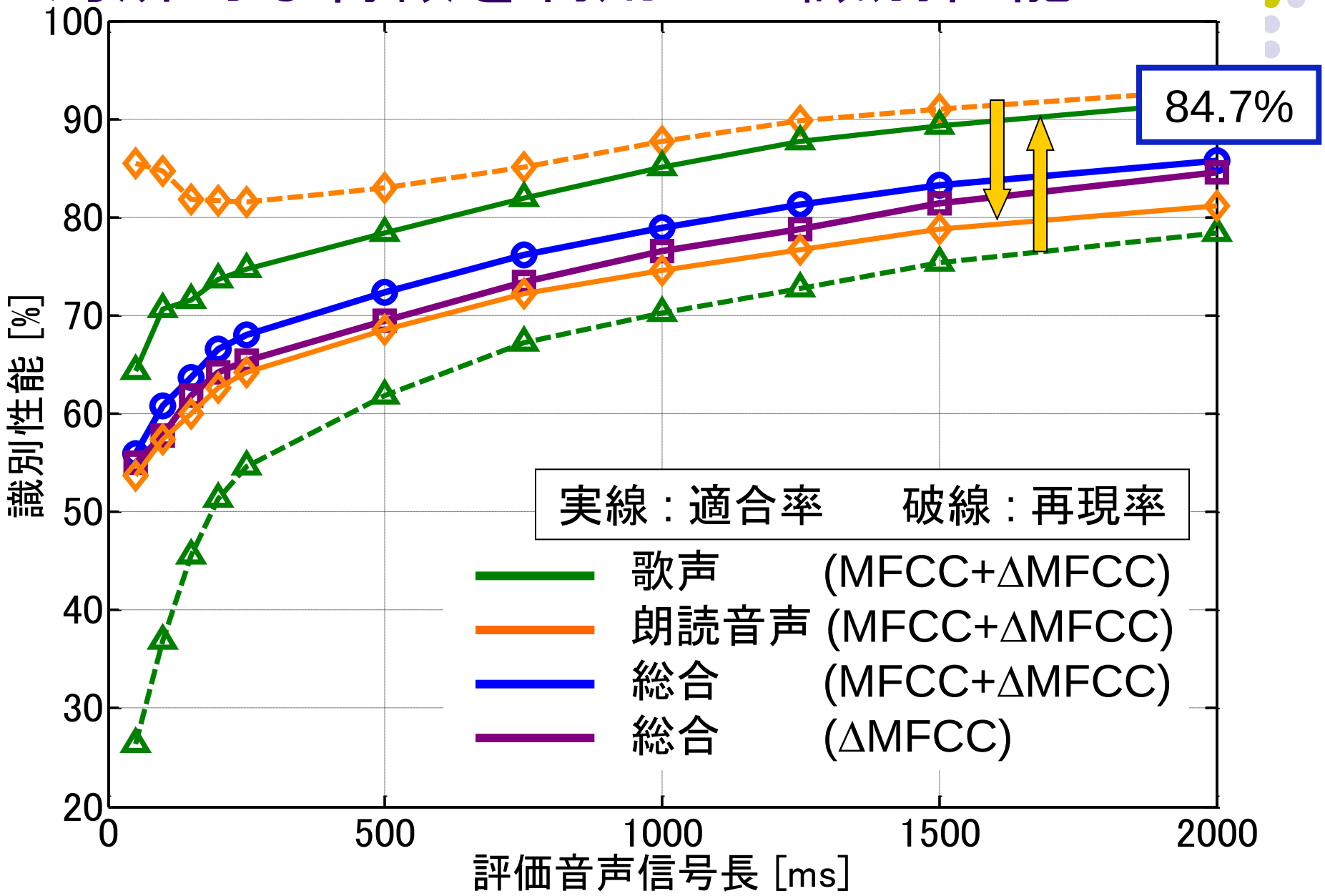
局所的な特徴を利用した識別性能



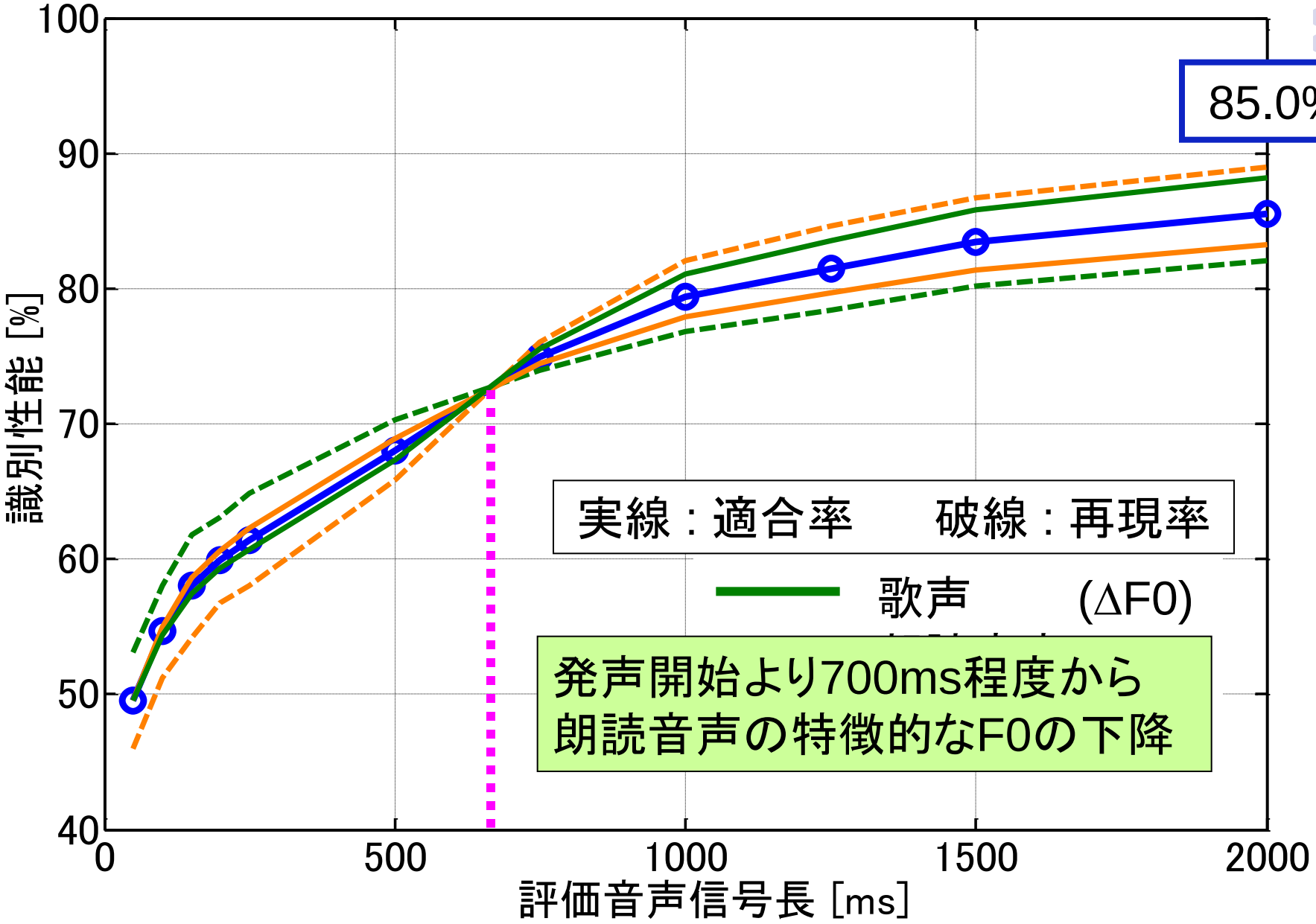
局所的な特徴を利用した識別性能



局所的な特徴を利用した識別性能

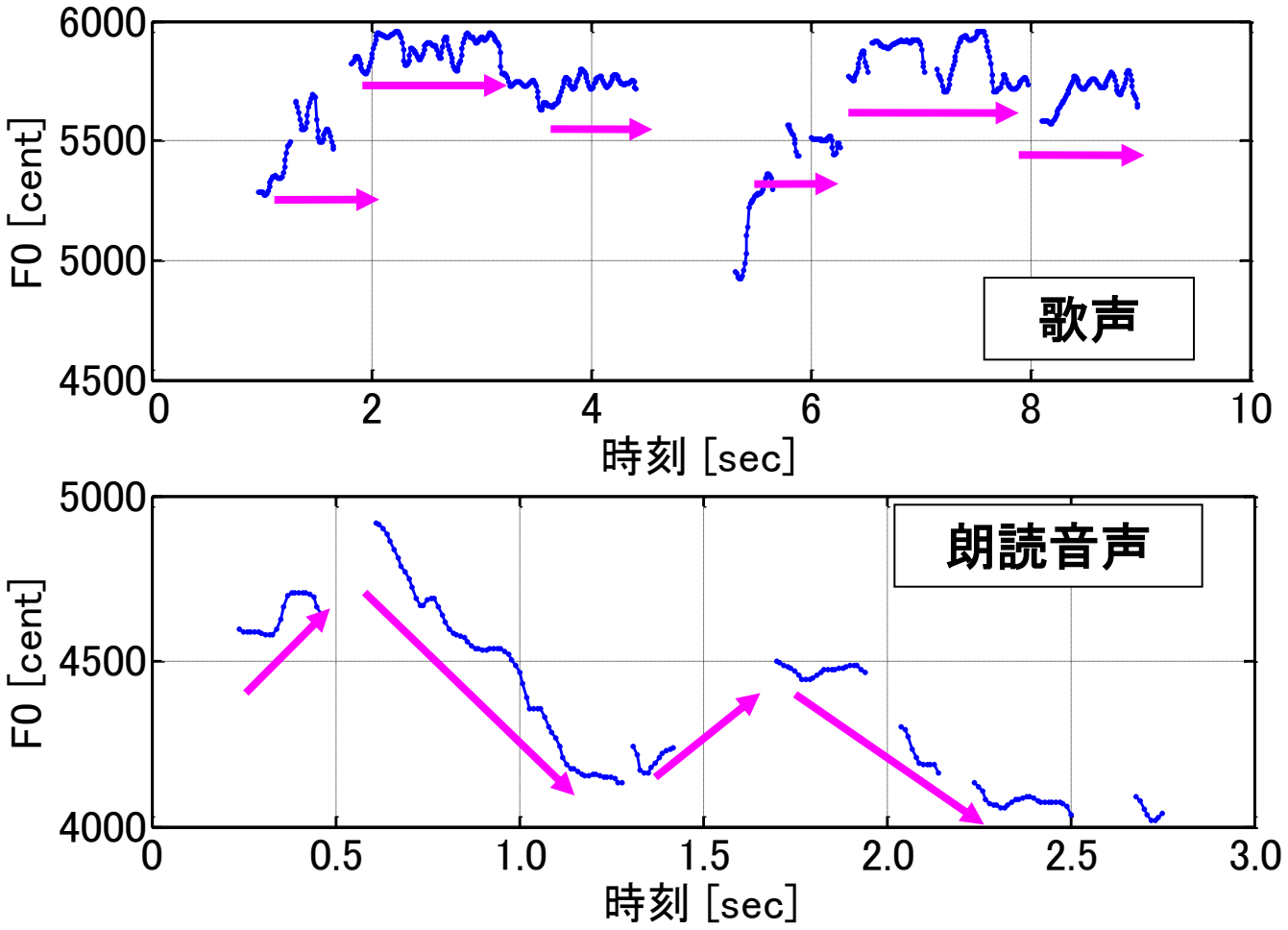


大局的な特徴を利用した識別性能



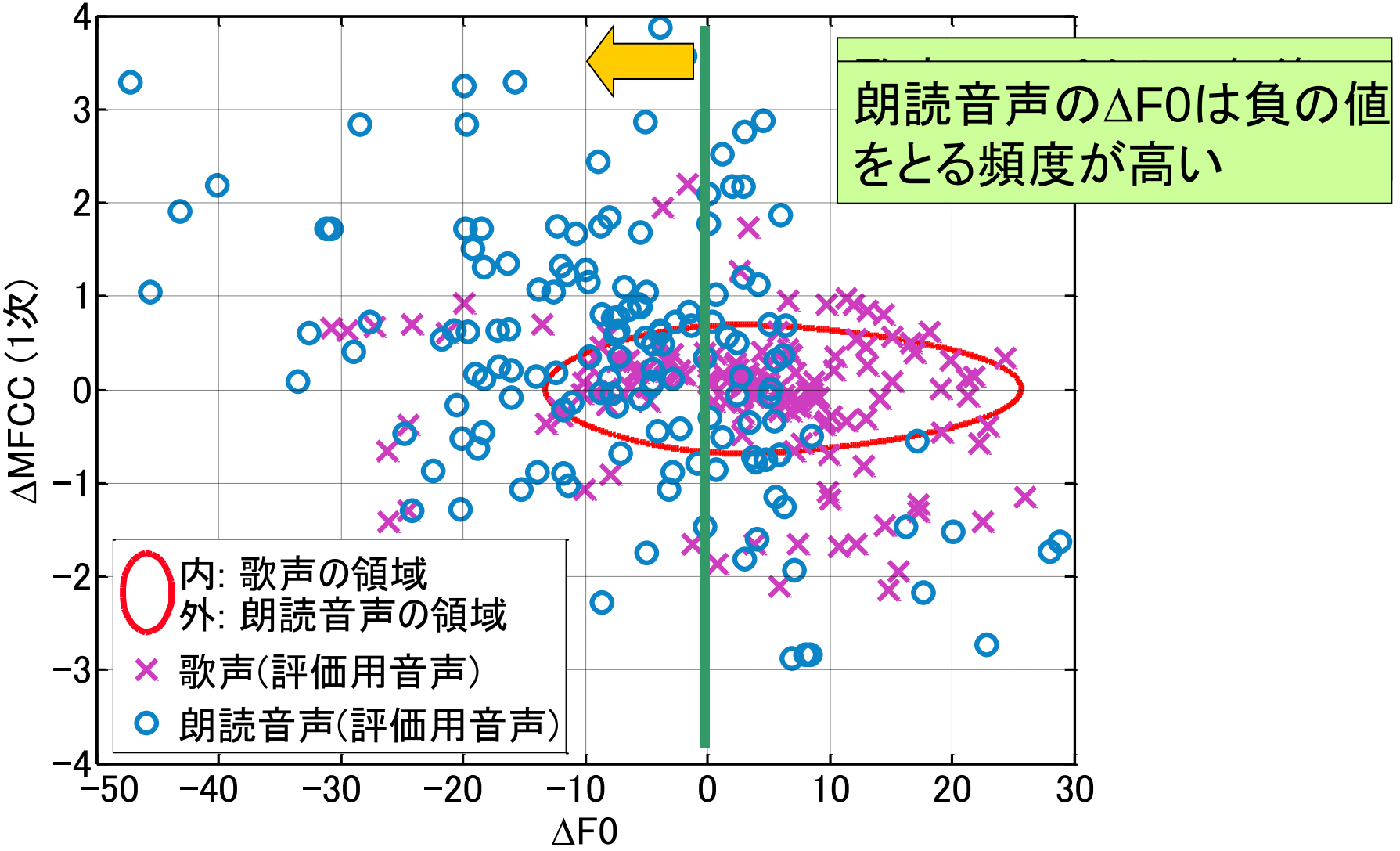
大局的な特徴を利用した識別性能

大局的特徴　音高の変化やリズム，テンポの違い

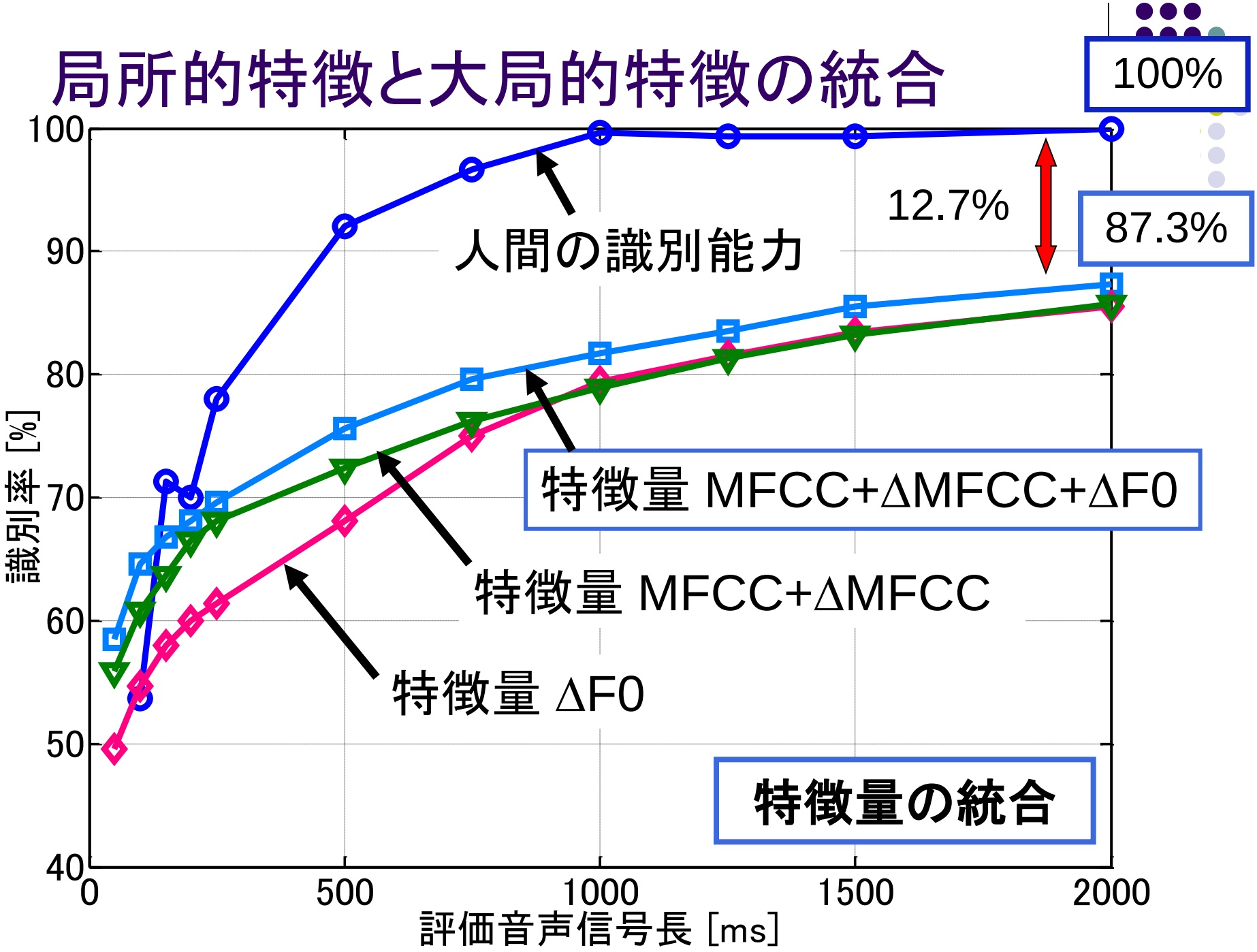


局所的特徴と大局的特徴の統合

- ΔF0とΔMFCCによる歌声，朗読音声の識別境界



局所的特徴と大局的特徴の統合



考察

- 短時間の音声信号の識別
  - MFCCとΔMFCCが有効

⇒スペクトル包絡とその時間変化の違い

- 1秒よりも長い音声信号の識別
  - ΔF0の利用

⇒F0の時間変化の違い

Δ算出の時間幅は50msと非常に短い

- 歌声：F0の時間変化が小さい
- 朗読音声：F0の下降する頻度が高い

F0の補間により長時間から算出されるΔF0の利用

まとめ

- 局所的･大局的な特徴を利用した
  歌声，朗読音声の自動識別手法の提案
- MFCCを利用した場合
  - 識別率 68.1% (250msの音声信号)
- ΔF0を利用した場合
  - 識別率 85.0% (2秒の音声信号)
- 2つの尺度の統合
  - 識別率 87.3% (2秒の音声信号)
- 人間の識別能力と比較
  - 提案手法の識別性能は低い

今後の展開

- 新たな識別特徴量の検討
  - 時間領域での音声信号の特徴量
    振幅やパワー, それらの時間変化
- 音声信号に変形・破壊を施すことによる
  識別能力の変化 (聴取実験)
  - 音声のどのような特徴が識別手がかりとなるのか?
- 複数の識別特徴量の統合方法の検討
- 時系列の変化をモデル化する識別手法
  - マルコフモデル