2 - Q - 9

# 基本周波数とスペクトル包絡を利用した歌声と朗読音声の識別に関する検討

◎大石康智　宮島千代美　西野隆典　伊藤克亘　武田一哉（名大）　後藤真孝（産業技術総合研究所）

## はじめに

### ■ 音声の状態

- 基本周波数
- スペクトル包絡
- 音韻の発声方法
- 呼気の状態

音声
- 朗読音声
- 歌声
- 鼻歌
- 自由発話
- 独り言
- 泣き声
- 囁き声
- 言いよどみ
- 笑い声

**目的：計算機による音声の状態の識別器**

- 自律型ロボットの音声に対する聴覚的情景分析
- 音声認識システムにおける発話区間検出
- 音声の特徴を生かした検索システム (歌声，言いよどみ)

第一歩として，

歌声とその歌詞を朗読した音声(朗読音声)との識別

## 歌声と朗読音声の識別手法

### ■ 歌声・朗読音声の基本周波数，スペクトル包絡の違いに注目

歌声の特徴は
- メロディ(基本周波数)が音名 (C,C#,…) に対応して変化
- メロディの制約により倍音構造が鮮明に現れるのではないか？
- 引き伸ばす発声が多いため，スペクトル包絡の変化が小さいのではないか？

### ■ 特徴量抽出

**基本周波数の抽出**

音楽のメロディの音高推定手法 (Goto,2004)
- 基本周波数
- 倍音パワー
  (第10倍音までのパワーの総和)

10msごとに抽出

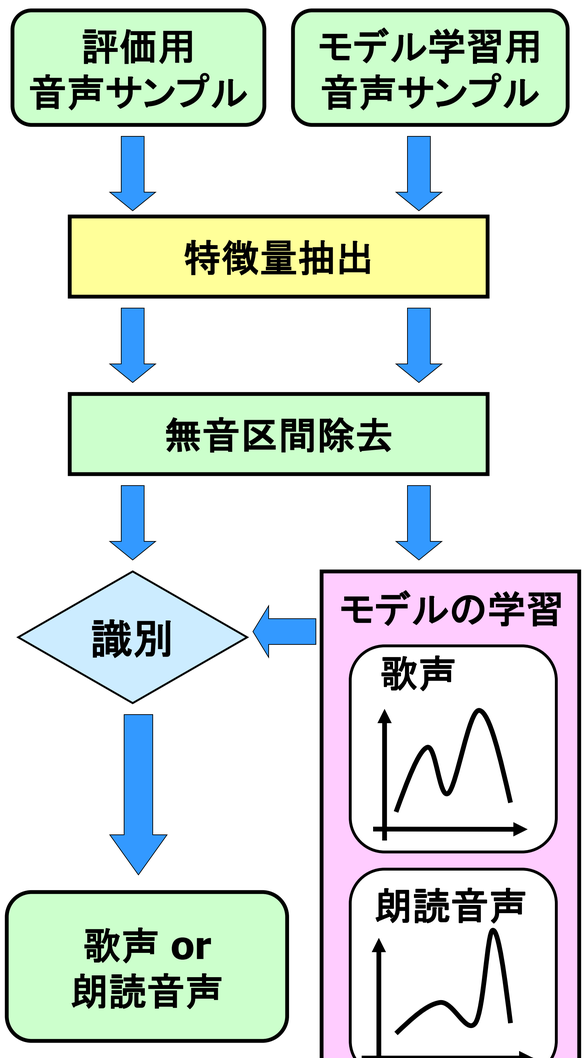
**MFCCの抽出**

HTK (Hidden Markov Model Toolkit)を利用

音声の分析条件

| | |
|---|---|
| 標本化周波数 | 16kHz |
| 分析窓 | ハミング窓 |
| フレーム長 | 100ms |
| フレームシフト | 10ms |
| 使用帯域 | 0～8000Hz |
| メルフィルタバンク数 | 24個 |
| MFCC次数 | 12次 |

### ■ 識別の流れ




### ■ モデルの学習と評価

**モデルの学習**

すべての学習用音声サンプルの特徴ベクトル系列

EMアルゴリズム

多次元混合ガウス分布(GMM)

2つのモデル $\{M_{歌声}, M_{朗読音声}\}$ を推定

**識別方法**

クロスバリデーション

学習用：評価用 = 3：1

$$\hat{d} = \arg\max_{d=歌声,\ 朗読音声} \sum_{t=1}^{N} \log p(\boldsymbol{o}_t | M_d)$$

特徴ベクトル系列 $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots \boldsymbol{o}_N)$ に対して最大事後確率となるモデル $\hat{d}$ を選択

## 評価実験

### ■ 使用データ

- 75名(男性38名，女性37名)の音声データベース

被験者1人あたりの収録データ

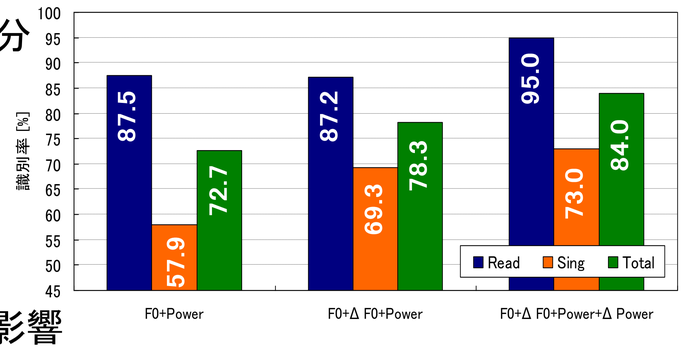
| | 歌声 | 朗読音声 |
|---|---|---|
| 出だし | 25サンプル | 25サンプル |
| サビ | 25サンプル | 25サンプル |

- RWC研究用音楽データベースから25曲を抜粋 (平均 119bpm)

**基本周波数と倍音パワーの利用**

- 基本周波数+倍音パワー+ Δ 成分

条件
GMMの混合数　4
Δ の時間幅　40ms



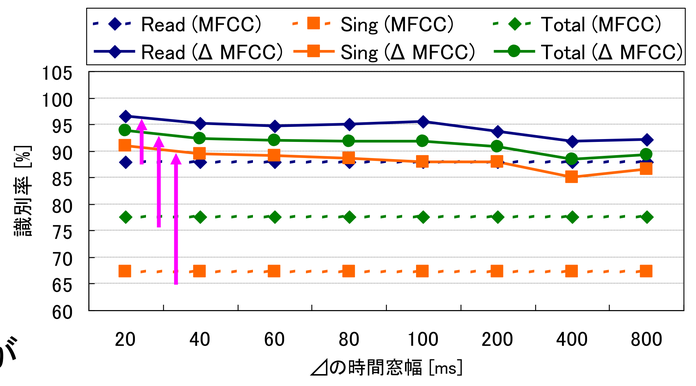

局所的な時間変化成分が識別に影響

→ **基本周波数，呼気の変化の違いをモデル化**

**ΔMFCCの利用**

- Δ MFCC12次

条件
GMMの混合数　16
MFCCのフレーム長　100ms

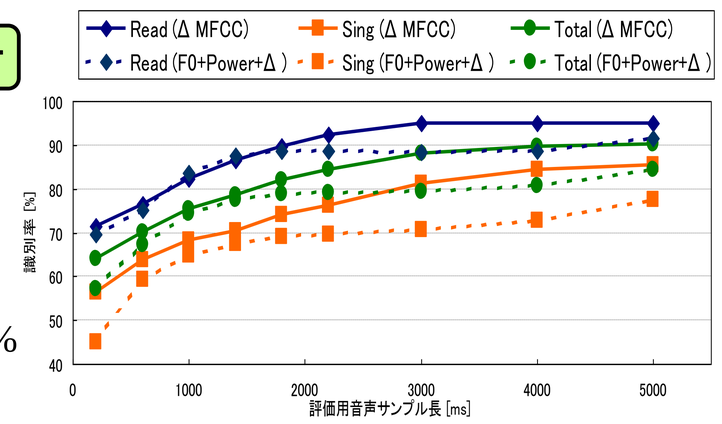



スペクトル包絡の時間変化成分が識別に影響

→ **歌声特有の伸ばす発声をモデル化**

**評価するサンプル長の検討**

条件
GMMの混合数　16
MFCCのフレーム長　100ms




5秒を分析したとき識別率 90.1%

有音区間：無音区間 = 4：1

→ **実行長約 6.25s，発声開始から3小節程度で識別可能**

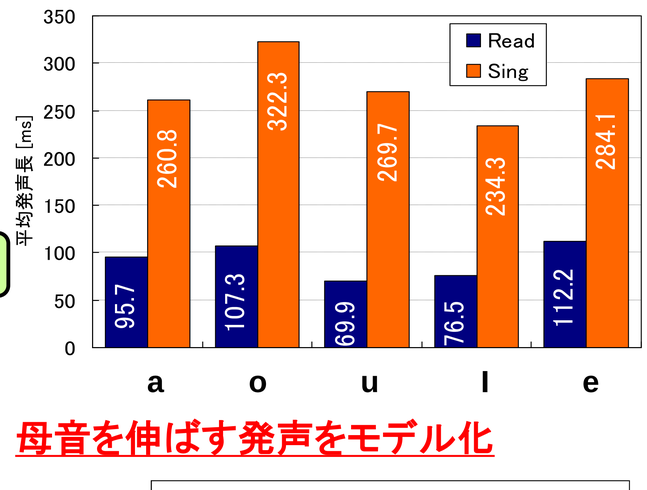
## 考察

**考察1:** 歌声と朗読音声の母音の発声長に違いはあるか？

**音声サンプルと歌詞のアライメント**

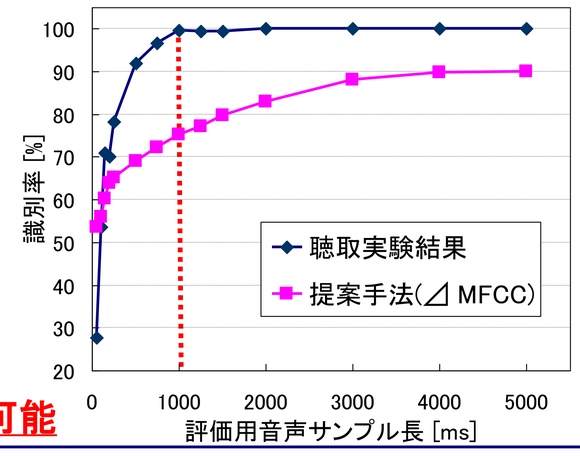
歌声は朗読音声に比べて母音は約3倍の長さ

**ΔMFCCの利用** → **母音を伸ばす発声をモデル化**




**考察2:** 人間はどの程度聴取すれば，歌声と朗読音声の識別が可能か？

**聴取実験**
- 音声サンプルの発声開始からある時間長で切り出す
- 被験者10名が，歌声か朗読音声か，判別不可かを回答




→ **人間は約1秒聴取すれば識別可能**

## まとめと今後の課題

- **基本周波数**と**スペクトル包絡の動的成分**を利用して，歌声と朗読音声の識別
  - ΔMFCCを使用して発声開始から4小節程度の観察 → **識別率：90.1%**
  - ↔ 聴取実験により人間は2拍程度の聴取で100%識別可能
- 発声開始の音階，リズムを考慮した特徴量の検討