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・情報科学), 後藤真孝 (産総研)

1 はじめに

歌声と朗読音声の識別について検討する．歌声に関する先行研究としては，歌声らしさの心理的要因に関する研究 [1]，歌声に自動で伴奏を付与する研究，歌声における基本周波数の動的変動成分 [2] を抽出したり，隠れマルコフモデルを利用することによって歌声を合成する研究 [3] がなされているが，歌声と話し声の識別に関する研究はまだ盛んに行なわれていない．歌声と話し声はどちらも音声のある一つの状態であると考えれば，そこには類似した特徴もあれば，異なる特徴もあると考えられる．その特徴から人間は，数秒聞いただけで，歌っているのか話しているのか判断が可能である．

本研究では，計算機による音声の状態の識別の第一歩として，歌声と朗読音声の識別を行なう．これは，音声に対する聴覚的情景分析の要素技術であり，将来，自律型ロボットとの対話などにおいても，音声の状態を識別することは大変重要な研究であると考えられる．そこで本報告では，歌声と朗読音声の基本周波数とスペクトル包絡情報に着目し，それらを特徴量として学習したパラメトリックなモデルによる識別手法について検討を行なう．

2 歌声と朗読音声の違い

歌声は，朗読音声と対比して，発声音高 (ピッチ) の幅 (音域) が広く，発声の強さの幅が大きいという特徴がある [1][2]．ここで，基本周波数の時間変化に着目すると，歌声特有の主な特徴として以下の 3 点が挙げられる．

1. 音名 (C, C#...) に対応して変化する傾向がある
2. ある音名に対応する定常部では，朗読音声のように周波数が下降せず，安定することが多い
3. ある音名から音名に変化する過渡部分において歌声特有の動的変動 (オーバーシュート・アンダーシュート，ヴィブラート，突発的変化) が現れる

また歌声では歌のメロディによる制約がかかるため，基本周波数のパワーが朗読音声に比べて強く，その倍音構造もはっきりと現れると考えられる．

一方，音声のスペクトル包絡に着目すると，歌声ではある音韻を引き延ばす発声が多いため，スペクトル包絡の変化が小さい傾向がある．

以上より本研究では，入力特徴量として音声の基本周波数とそのパワー，スペクトル包絡情報を持つ MFCC だけでなく，それぞれの時間変化量も利用する．

3 歌声と朗読音声の識別実験

3.1 使用データ

産業技術総合研究所によって収録された 75 名 (男性 38 名，女性 37 名) の被験者による音声データベースを使用した．被験者 1 人あたりの収録データは，RWC 研究用音楽データベース (RWC-MDB-P-2001)[4] から抜粋した 25 曲のポピュラー音楽に対して，各曲の歌の出だしの部分とサビの部分の歌詞を歌った 50 サンプルと，その歌詞を朗読した 50 サンプルで構成される計 100 サンプルである．

3.2 識別のための特徴量抽出

3.2.1 基本周波数

音楽のメロディの音高推定手法 [5] を利用して基本周波数とその第 10 倍音までのパワーの総和 (以後，パワーと呼ぶ) を 10ms ごとに算出した．

3.2.2 MFCC

MFCC(Mel-Frequency Cepstrum Coefficient) を算出する．分析条件としては標本化周波数 16kHz，分析窓はハミング窓で行なう．フレーム長は 25ms〜800ms の間で変化をさせ，フレームシフト幅は 10ms である．使用帯域は 0〜8000Hz でメルフィルタバンクを 24 個配置し，12 次までの係数を利用する．

3.2.3 Δ 成分の算出

基本周波数とそのパワー，MFCC の Δ 成分の算出式を以下に示す．ここで l はフレーム番号，K は回帰係数を求める時間幅である．

$$\Delta c(l) = \frac{\sum_{k=-K}^{K} kc(l+k)}{\sum_{k=-K}^{K} k^2} \tag{1}$$

3.2.4 無音区間の除去

無音区間の量により歌声と朗読音声が識別されるのを防ぐため，音声のパワーが弱い，出だし，終わり，休止の部分において算出された特徴量 (基本周波数 orMFCC) を除去する．

3.3 モデルの学習

識別のためのモデルとして，多次元混合ガウス分布 (GMM) を用いる．これは，音声認識で HMM の各状態の特徴ベクトルの確率密度関数を推定する際によく利用され，GMM に基づく識別器は，学習データ量に応じて比較的簡単にパラメトリックなモデルが推定できる特徴を持つ．歌声，朗読音声に対応する二つのモデルを構築するために，音声データベースの各サンプルから抽出した特徴量ベクトルを用いて，ガウス分布の重み，特徴ベクトルの平均，分散を EM アルゴリズムで推定する．ここでは対角共分散のみを推定する．

3.4 モデルの評価

評価方法としては使用データを 4 つのグループに分け，そのうちの 3 つを識別モデルの学習用，残りの 1 つを評価用とし，これを交叉させて 4 通りの評価実験を行うクロスバリデーションを採用した．

GMM により学習されたモデル $\{M_{\text{歌声}}, M_{\text{朗読音声}}\}$ に対して評価用サンプルの特徴量ベクトル系列 $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_N)$ が与えられたときの尤度は次式で定義する．与えられた特徴量ベクトル系列 $\boldsymbol{O}$ に対して最大事後確率となるモデル $\hat{d}$ を識別結果とする．

$$\hat{d} = \underset{d=\text{歌声, 朗読音声}}{\operatorname{argmax}} \sum_{t=1}^{N} \log p(\boldsymbol{o}_t | M_d) \tag{2}$$

以下に各特徴量とその Δ 成分を使用したときの識別率の変化を示す．混合ガウス分布の混合数は 4 混合，基本周波数とパワーの Δ 成分を求める時間幅は 40ms，MFCC の Δ 成分を求める時間幅は 100ms と固定した．

3.4.1 基本周波数とパワーの利用

図 1 より，基本周波数とそのパワーの Δ 成分の併用が識別率に大きく影響していることがわかる．特に基本周波数の Δ 成分の併用により，歌声の識別率は 11.4 ポイント向上したことから，歌声にはメロディの制約による基本周波数の変化に特徴があると考えられる．またパワーの Δ 成分を併用することで，朗読音声も歌声も識別率が上昇した．パワーの時間的な変化，つまり人間が歌うときと朗読するときの呼気の状態に大きな違いがあると考えられる．

---

*A sung speech and read speech discriminator based on fundamental frequency and spectral envelope by Y. Ohishi, C. Miyajima, T. Nishino, K. Ito and K. Takeda (Nagoya Univ.), M. Goto (AIST)

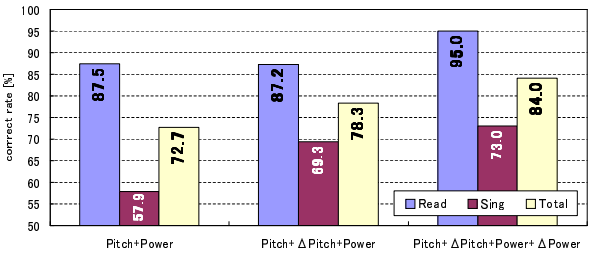

図 1: 基本周波数とパワーの Δ 成分を用いたときの識別率の変化

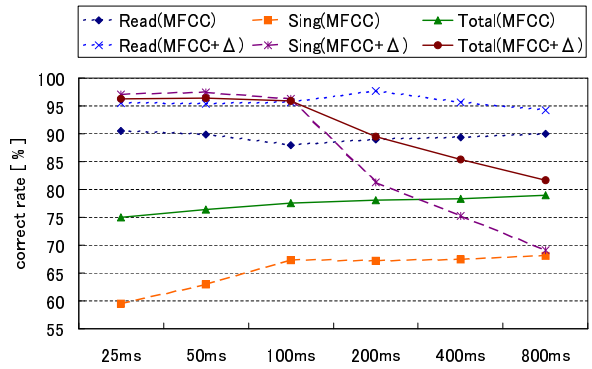

図 2: MFCC 算出のフレーム長の変化に対する識別率の変化

3.4.2 MFCC の利用

図 2 は MFCC を算出するフレーム長を 25ms〜800ms の間で変化させたときの識別率の変化である．MFCC だけを特徴量としたとき，歌声の識別率はフレーム長の増加とともに上昇した．全体の識別率は 25ms から 800ms の間で 4.0 ポイント向上した．これは歌声と朗読音声の音節の速度の違いをモデル化した結果であると考えられる．フレーム長を変化させることで歌声と朗読音声のスペクトル包絡にどのような変化が現れるかについては今後，特徴量を LPC から抽出することによって，より細かく検討する予定である．

また MFCC の Δ 成分を併用することにより，フレーム長 50ms のときの識別率は，MFCC のみを特徴量に利用したときに比べて，朗読音声 5.5 ポイント，歌声 34.5 ポイント上昇し，全体の識別率は 96.4%である．つまり，MFCC の Δ 成分によって識別率が上昇したということは，歌声独特の音韻を引き伸ばすという発声の仕方をモデル化できたと考えられる．フレーム長が長い場合の識別率の低下に関しては，そのフレーム長に適した Δ 成分算出の時間幅について検討する必要がある．

図 3 は歌声と朗読音声の MFCC のモデルにおける平均値から逆離散コサイン変換を行なうことにより算出した 24 個のフィルタバンクの出力から推定されるスペクトル包絡である．低域の部分の歌声の包絡には，朗読音声には存在しない倍音構造のような外形が見られる (図 3 の破線丸印)．これは，歌声がメロディの制約により，倍音構造が鮮明に現れるという傾向と整合性がとれる．

また，以上より MFCC で歌声をモデル化するということは，基本周波数の倍音構造も含めてモデル化していることになるため，基本周波数とそのパワーでの識別率よりも MFCC のモデル化による識別率のほうが性能が良いと考えられる．

3.4.3 評価するサンプル長の検討

評価用サンプルの各モデルに対する事後確率を求めるサンプル長 N (式 (2)) を変化させ，その識別率について検討を行なう．図 4 は，識別に利用するサンプル長 N を増やしたときの識別率の推移である．このとき MFCC を算出するフレーム長は 50ms とした．

特徴量として MFCC とその Δ 成分を用いた場合，サンプル長が 200ms では全体でも 71.2%の識別率が，サンプル長が 5s では全体で約 94.1%の識別率が得られ，20 ポイント以上の向上が見られる．

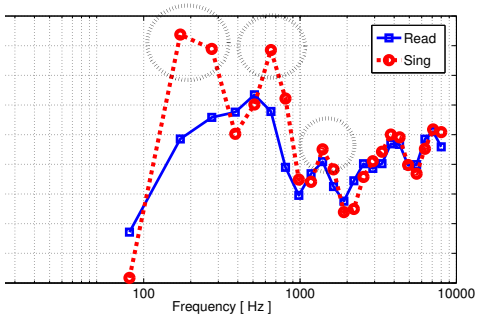

図 3: MFCC の平均値から算出したスペクトル包絡

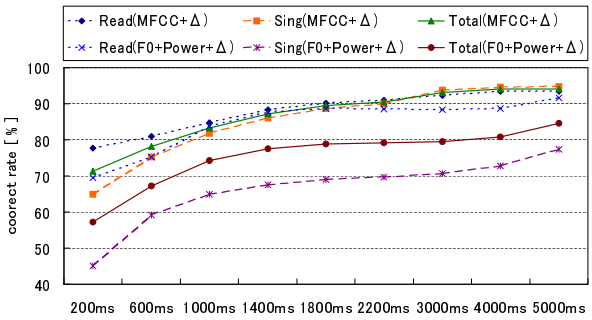

図 4: 評価するサンプル長を変化させたときの識別率の変化

今回使用した歌声データの平均 BPM(Beat Per Minute) は 119 であった．また，1 サンプルあたりの有音区間と無音区間の比率は約 4 : 1 であることから評価用音声サンプルの実行長約 6.25s，すなわち曲の歌い始めから 4 小節程度観察することで，歌声か朗読音声かの識別が可能であるということが確認できる．

4 まとめと今後の展開

本報告では歌声と朗読音声の識別を行なった．識別特徴量は基本周波数とそのパワー，MFCC を使用し，各特徴量の Δ 成分を使用することにより，識別率の大幅な向上がみられる．これは呼気の状態，音韻の発声方法，倍音構造の違いをモデル化したことによるものである．最も高い識別率は MFCC とその Δ 成分を使用したときで，96.4%であった．今後の展開としては，フレーム長に応じた Δ 成分の算出時間幅，特徴量の次元数の最適化，さらに有効な特徴量抽出の検討が必要である．また，話し言葉や笑い声などその他の音声の状態との識別についても検討していく予定である．

参考文献

[1] 辻直也, 赤木正人, ”歌声らしさの要因とそれに関連する音響特徴量の検討”, 日本音響学会聴覚研究会資料,pp41-46,Vol.34,No.1,H-2004-8.

[2] 斉藤毅, 鵜木祐史, 赤木正人, ”歌声の F0 制御モデルにおけるパラメータ決定に関する考察”, 日本音響学会聴覚研究会資料,pp653-658,Vol.33,No.10,H-2003-111.

[3] 酒向慎司, 宮島千代美, 徳田恵一, 北村正, ”隠れマルコフモデルに基づいた歌声合成システム”, 情報処理学会論文誌,pp719-727,Vol.45,No.3, March 2004.

[4] 後藤 真孝, 橋口 博樹, 西村 拓一, 岡 隆一: ”RWC 研究用音楽データベース: 研究目的で利用可能な著作権処理済み楽曲・楽器音データベース”, 情報処理学会論文誌, Vol.45, No.3, pp.728-738, March 2004.

[5] Masataka Goto, A Real-time Music-scene-description System: Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-world Audio Signals, Speech Communication, Vol.43, No.4, pp.311-329, 2004.