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† 　後藤　真孝†† 　伊藤　克亘† 　武田　一哉†

† 名古屋大学大学院情報科学研究科
〒 464-8603 愛知県名古屋市千種区不老町
†† 産業技術総合研究所
〒 305-8568 茨城県つくば市梅園 1-1-1
E-mail: †ohishi@sp.m.is.nagoya-u.ac.jp, {k-ito, takeda}@is.nagoya-u.ac.jp,
††m.goto@aist.go.jp

あらまし　音声信号の局所的・大局的な特徴を利用した歌声と朗読音声の識別について検討する．聴取実験の結果, 人間は 200ms, 1s の音声信号に対して, それぞれ 70.0%, 99.7%で歌声と朗読音声の識別が可能であることを確認した．この結果より, 短時間・長時間の音声信号に対して, 異なる特徴が識別に影響するということを想定し, スペクトル包絡 (MFCC) と基本周波数の軌跡の 2 つの尺度に基づく識別器を設計した．このとき, 入力音声信号が 1 秒よりも長い場合, 基本周波数の軌跡を特徴量として利用した方がスペクトル包絡を特徴量とするよりも識別性能が高い．特に, 発声開始より 2 秒の音声信号に対して 85.0%の歌声と朗読音声の識別が可能であった．一方, 入力音声信号が 1 秒よりも短い場合, スペクトル包絡の方が基本周波数の軌跡に比べて識別性能が高い．最終的に, 2 つの尺度を単純に統合することによって 2 秒の音声信号に対して 87.5%の識別率を得ることができた．

キーワード　歌声, 朗読音声, 音声の識別, スペクトル包絡, 基本周波数, 混合ガウス分布

Discrimination between Singing and Speaking Voices Using Local and Global Characteristics

Yasunori OHISHI†, Masataka GOTO††, Katunobu ITOU†, and Kazuya TAKEDA†

† Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8603, Japan
†† National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan
E-mail: †ohishi@sp.m.is.nagoya-u.ac.jp, {k-ito, takeda}@is.nagoya-u.ac.jp,
††m.goto@aist.go.jp

Abstract　Discriminating between singing and speaking voices by using the local and global characteristics of voice signals is discussed. From the results of subjective experiments, we show that human beings can discriminate singing and speaking voices with more than 70.0% and 99.7% accuracy from 200 ms and one second long signals, respectively. From the subjective experiment results, assuming that different features are effective for short-term and long-term signals, we designed two measures using a spectral envelope (MFCC) and the fundamental frequency (F0, perceived as pitch) contour. Experimental results show that the F0 measure performs better than the spectral envelope measure when the input voice signals are longer than one second. Particularly, it can discriminate singing and speaking voices with 85.0% accuracy with two-second signals. On the other hand, when the input signals are shorter than one second, the spectral envelope measure performs better than the F0 measure. Finally, by simply combining the two measures, 87.5% accuracy is obtained for two-second signals.

Key words　Singing Voice, Speaking Voice, Voice Discrimination, MFCC, F0, GMM

1. はじめに

人間の口から発する音には, 話し声, 歌声, 笑い声, 咳, 囁き声, リップノイズのようにさまざまな音響的事象がある. 人間は, これらの事象を上手に使い分けることによって複数の相手とのコミュニケーションを成り立たせている. それは, 人間が瞬時に音を理解し, 自動的に識別することが可能だからである. 我々はこの人間による音の識別を理解することによって, 計算機上での音の自動識別を目指している.

本研究では, これらの音響的事象の中の歌声と朗読音声の識別に着目する. 歌声と朗読音声の違いについては多くの研究がなされており, 歌声の典型的な特徴としては基本周波数 (以後, F0 と呼ぶ) とその強度が幅広く変化し, スペクトル包絡に関して言えば, 歌声は *SingingFormant* と呼ばれる特別なフォルマントが存在する [3], [5]〜[8], [11], [12], [15]. ただこの *SingingFormant* は, オペラ歌手の歌声から観測されたものである. これは喉頭の部分で共鳴を起こし, 深い響きを作り出す洋楽的歌唱法とされており, 必ずしも素人の歌声に観測できるとは限らない. しかし, 人間はたとえ歌唱者が素人であったとしても, 日常会話の話し声との識別が可能である. それは歌声の性質ばかりでなく, 歌い方と話し方の違いを人間は識別しているのではないかと考えられる.

近年, 様々な音楽と音声のカテゴリの識別手法が数多く提案されてきた [2], [9], [10]. それらの手法を, 歌声と朗読音声の識別に適用することは困難である. なぜなら音楽のカテゴリとして楽器音のみや伴奏付きの歌声が対象であったために, 混合音の特徴量が主に検討されていたからである. つまり, 伴奏のない歌声そのものの特徴は, まだ十分に議論されていなかった.

本研究の目的は歌声に含まれる物理的な性質を特徴づけ, 話し声との違いを明らかにすること, また歌い方, 話し方というように長時間から観測できる発声のスタイルの違いを識別するための尺度を構築することである. 以下, 2 節では聴取実験に基づいて歌声と朗読音声の人間の識別能力を調べ, 1s の音声信号に対して 99.7%の識別率で人間が識別できることを示す. 次に 3 節では, 歌声と朗読音声とを計算機が自動識別するための具体的な手法として, 局所的な特徴としてスペクトル包絡, 大局的な特徴として F0 の軌跡の二つの特徴を用いた識別手法を提案する. 4 節では使用した歌声データベースについて述べ, 5 節では評価実験の結果を示す. 最後に, 6 節で実験結果に対する考察を述べ, 7 節で本研究のまとめを行う.

2. 歌声と朗読音声の人間の識別能力

歌声と朗読音声は, どの程度の音声信号長を聴取すれば識別が可能であるか調査した. ここでは 4 節で説明する音声データベースから女性 25 名, 男性 25 名を選び, 25 曲の歌声, またはその歌の歌詞を朗読している朗読音声を用いて, 発声開始から 10 段階の異なる長さで切り出したもの 50,000 サンプルを使って聴取実験を行った. まず, この音声信号 50,000 サンプルの中から, 音声信号 500 サンプル (歌声 250 サンプルと朗読音声 250 サンプル) を表 1 のように切り出した長さごとにランダムに選

表 1　聴取実験の評価セット

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	250 サンプル	250 サンプル

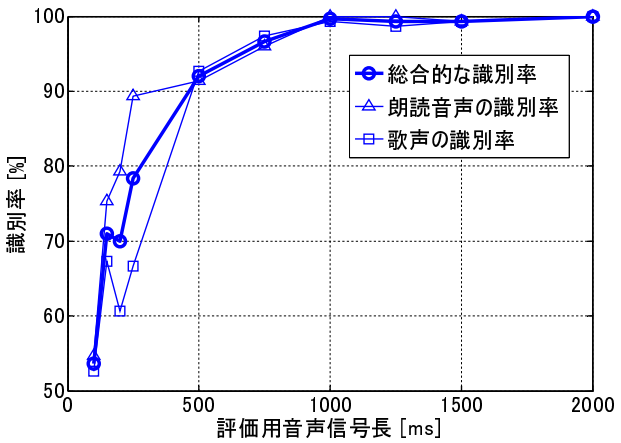

図 1　歌声と朗読音声を人間が聴取して判断する場合の識別率: 歌声の識別率は歌声を聴取したとき歌声と正答した割合, 朗読音声の識別率は朗読音声を聴取したとき朗読音声と正答した割合, 総合的な識別率は両者の平均値である.

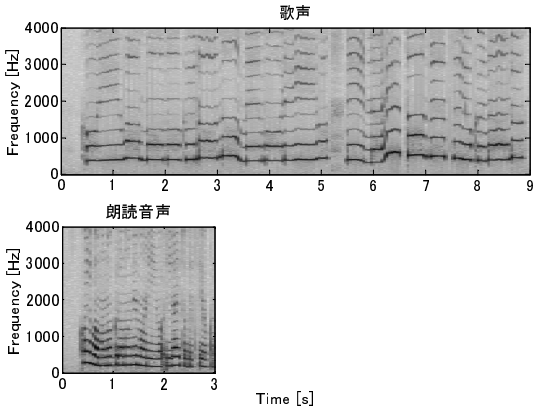

図 2　歌声と朗読音声のスペクトログラム（同一歌詞を発声）

んだ評価セットを 10 種類作る. 10 人の被験者ごとに異なる評価セットを割り当て, その全サンプルをランダムな順番で 1 回だけ聴取させ, 各サンプルが, 「歌声」であるか, 「朗読音声」であるか, もしくはあまりに聴取時間が短いため「識別不可能」かの 3 通りで回答させた.

また, 聴取実験後に, 被験者に感想及び識別の判断基準について質問した結果, 以下のような意見を得た.

- 最低 1 音節必要
- 音高の変化の違いに注目
- リズム, 発声の長さの違いに注目
- 1 音節目の母音が長いと歌声ではないか?
- 歌い始めと, 朗読し始めの息の吸い方が異なる

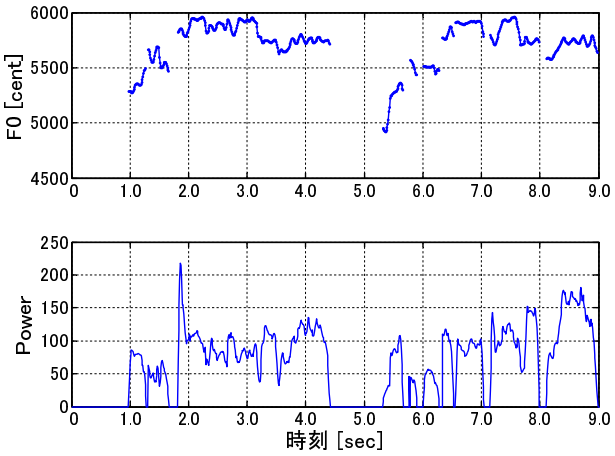

図 3　歌声の F0 の軌跡とパワー

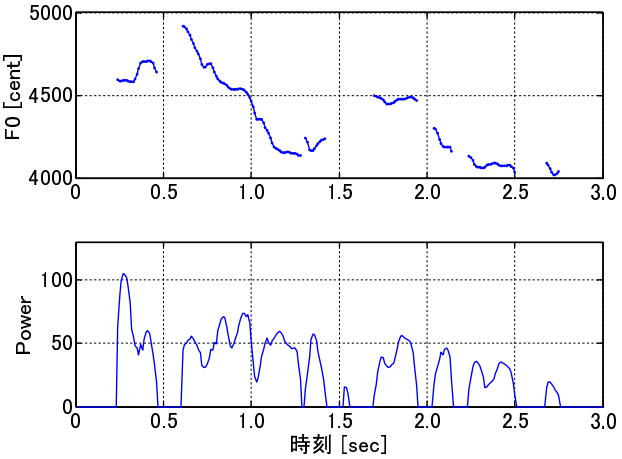

図 4　朗読音声の F0 の軌跡とパワー

- 1 音節目の呼気の量の違いに注目

以上より, 被験者は様々な観点に基づいて音声の識別を行っていると考えられる.

図 1 に示されるように, およそ発声開始から 1 秒程度の音声信号の聴取により, 人間は歌声か朗読音声かの識別が可能であることがわかる 200ms の時点で既に識別率は 70.0%を超えており, 特に短時間の場合, 朗読音声の識別率が高い傾向がある. このことは歌声のリズムやメロディに対応する大局的な特徴だけでなく, スペクトル包絡のような短時間の局所的な特徴も, 識別の手がかりになっているのではないかと考えられる. これらの観察に基づいて, 以下の節では歌声と朗読音声の識別に関して, 2 つの尺度を検討する.

3. 識別尺度

本節では, 歌声と朗読音声を識別するにあたって, 2 つの異なる尺度, すなわち音声信号から局所的・大局的に観測される特徴を利用する. 具体的には, 局所的な特徴としてはスペクトル包絡を, 大局的な特徴として音声信号から抽出される F0 の軌跡を利用する.

3.1　音声信号の局所的な特徴に基づく尺度

歌声は, 2000Hz あたりに朗読音声には見られない *Singing-Formant* と呼ばれる加法的な共振をもつ [5]. また, 掠れた声のスペクトルの外形は朗読音声よりも傾斜が急斜である [1]. それゆえに, 短時間の音声信号から抽出されるスペクトル包絡は音声の識別をするためのひとつの手がかりになると考えられる.

図 2 は, 歌声と朗読音声のスペクトログラムである. これは発声者がある歌を歌い, またその歌詞を朗読したときのものである. このとき歌声の方が広帯域にわたって倍音構造が鮮明に現れ, スペクトル包絡の違いが観察できる.

スペクトル包絡の尺度として, メル周波数ケプストラム係数 (MFCC) とその時間変化成分 (ΔMFCC) を利用する. この特徴量は音声認識システムにおけるスペクトル包絡の抽出に利用されている. 分析条件は以下のとおりである. ΔMFCC は, (式 1) のように $2K+1$ 個のフレームにわたる回帰係数を計算した.

表 2　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	25ms
フレームシフト	10ms
メルフィルタバンク数	24
使用帯域	0〜8000Hz

$$\Delta c[n] = \frac{\sum_{k=-K}^{K} k \cdot c[n+k]}{\sum_{k=-K}^{K} k^2} \qquad (1)$$

歌声, 朗読音声それぞれの MFCC, ΔMFCC ベクトルの分布を 16 混合ガウス分布でモデル化する. 混合ガウス分布の共分散行列は, 対角共分散行列とする. 識別方法は, 以下のように最大事後確率に基づいて識別した.

$$\hat{d} = \underset{d=\text{歌声, 朗読音声}}{\operatorname{argmax}} f(\mathbf{x}; \Lambda_d) \qquad (2)$$

ここで Λ_d, (d = 歌声, 朗読音声) は MFCC ベクトルの分布に対する GMM のパラメータである.

3.2　音声信号の大局的な特徴に基づく尺度

歌声は曲のメロディとリズムパターンの制約を受けて生成されるため F0 の遷移が朗読音声とは異なると考えられる. それゆえに音声信号から抽出される F0 の遷移の違いを捉えることは, 歌声と朗読音声の識別のための手がかりになると考えられる. よってこの特徴を取り込むために, F0 の軌跡の時間変化 ΔF0 を利用する.

3.2.1　F0 抽出

F0 は, 後藤ら [13] の提案した優勢な F0 推定手法を利用して推定した. この手法は, 非周期的な雑音に加え, 高調波構造をもつ弱い雑音も含まれる場合を考慮して入力音声信号中で最も優勢な (パワーの大きい) 高調波構造の基本周波数を, 音声の F0 として抽出する. そのために, コムフィルタの考え方に基づいたフィルタを用いて, 時刻 t において周波数 F が F0 となる

$P_{F0}(F,t)$ を評価する．この手法を利用して, F0 を 10ms ごとに算出した．

次に以下のように [Hz] で与えられる周波数の単位を [cent] に変換する．

$$f_{\mathrm{cent}} = 1200 \log_2 \frac{f_{\mathrm{Hz}}}{440 \times 2^{\frac{3}{12}-5}} \qquad (3)$$

3.2.2 ΔF0 の分布

日本語の朗読音声の韻律は, 下降する F0 の軌跡によって特徴づけられる．そこで, ある時間幅にわたって計算された ΔF0 の分布を歌声と朗読音声の識別に利用する．

まず, 図 3, 4 のように F0 の軌跡とパワーの変化に着目する．発声区間ごとの F0 を, パワーが無音の箇所を除去して切り出し, 各発声区間で 5 点の回帰係数 (式 1 で $K = 2$) を求めることによって ΔF0 を計算する．ΔF0 の分布は, 歌声, 朗読音声ともに 16 混合ガウス分布 (対角共分散を利用) によってモデル化した．以下のように事後確率を最大にする音声信号を識別結果とする．

$$\hat{d} = \underset{d=\text{歌声, 朗読音声}}{\operatorname{argmax}} f(\mathbf{y};\Omega_d) \qquad (4)$$

ここで Ω_d, ($d =$ 歌声, 朗読音声) は ΔF0 の分布をモデル化する GMM のパラメータである．

4. 歌声データベース

本研究では, 産業技術総合研究所 (AIST) によって収録された歌声研究用音楽データベース「AIST ハミングデータベース」[14] の一部である, 日本人歌唱者 75 名分（男性 37 名, 女性 38 名）の音声データを抜粋して使用した．

各歌唱者が, *"RWC Music Database: Popular Music"* (RWC-MDB-P-2001) [4] 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, またその歌詞を読み上げた音声を用いた．つまり 1 名あたり計 100 サンプル (歌声: 50 サンプル, 朗読音声: 50 サンプル) となり, 75 名全員で 7500 サンプルとなる．音声サンプルの長さの平均は歌声で約 8 秒程度, 朗読音声で約 5 秒程度であった．

5. 提案手法の評価

本節では, まず音声信号の局所的な特徴であるスペクトル包絡を利用した歌声と朗読音声の識別手法について評価する．次に大局的な特徴, すなわち ΔF0 を利用した識別手法を評価する．最後に局所的な特徴と大局的な特徴についての識別性能の比較し, それらを組み合わせた手法も評価する．

5.1 評価方法

男性 37 名を M1〜M37, 女性 38 名を F1〜F38, 25 曲の歌の出だし部分を D1〜D25, 歌のサビ部分を S1〜S25 と記述したとき, 表 3 のように 15 個のグループを作成した．例えば, グループ A1 は, 男性 13 名, 女性 12 名の 10 フレーズからなる歌声 250 サンプルと朗読音声が 250 サンプル含まれていることになる．ここで A1 で評価を行う場合は, B2, B3, B4, B5, C2, C3, C4, C5 でモデルを学習する．つまり, 評価する音声信号のグループに含まれていない「発声者」, 「楽曲」でモデルを学習し, 評価を行う．これを評価対象グループを変えながら 15 回のクロスバリデーションを行い, その識別性能の平均値を識別率とする．

表 3　識別実験の学習, 評価セットの作成

	D1-10	S1-10	D11-20	S11-15 S21-25	D21-25 S16-20
M1-M13, F1-F12	A1	A2	A3	A4	A5
M14-M26, F13-F26	B1	B2	B3	B4	B5
M27-M37, F27-F38	C1	C2	C3	C4	C5

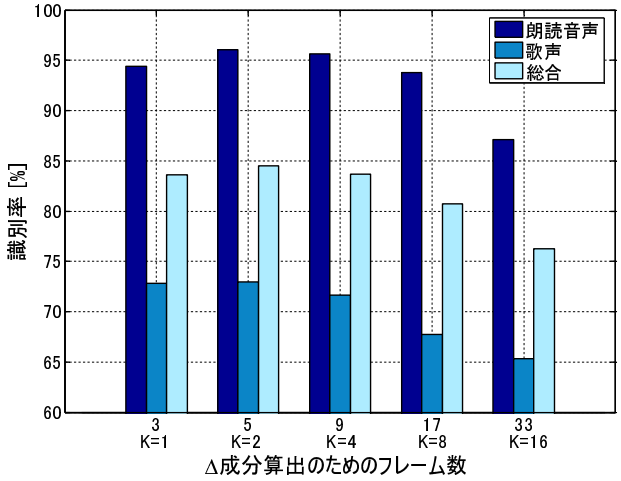

図 5　ΔMFCC を算出するフレーム数に対する識別率

5.2 局所的な特徴 (MFCC) を利用した識別性能

スペクトル包絡を利用した歌声・朗読音声の識別性能を評価する．MFCC は 12 次までの係数を利用する．

まず MFCC の時間変化として ΔMFCC を計算するための最適な時間幅 (式 1 のフレーム数 K) に関する検討を行った．図 5 は, ΔMFCC を算出するフレーム数を変化させたときの識別率の推移である．評価に利用した音声信号長は各サンプルの発声開始から 2 秒間である．このとき, ある時刻 t のフレームに対して前後 2 フレーム ($K = 2$), すなわち 5 フレームから算出した ΔMFCC の 12 次元ベクトルによって最も高い識別率が得られた．以後, ΔMFCC を算出する時間幅は前後 2 フレームから算出することにする．

図 6 は, MFCC による歌声と朗読音声の識別と ΔMFCC による識別との比較である．総合 (平均) 識別率が, 評価音声の時間長が長くなるにつれて単調に上昇していくことがわかる．MFCC を利用した場合, 2 秒間の評価音声に対して 72.4%の識別が可能である．このことから歌声と朗読音声のスペクトル包絡に違いがあるのではないかと考えられる．

一方, ΔMFCC を利用した場合, 2 秒の評価音声に対して 84.5%の識別が可能であり, MFCC の識別率を 12.1%上回った．このことからスペクトル包絡の局所的な変化にも歌声と朗読音声の違いがあると考えられる．

図 7 は, MFCC と ΔMFCC を連結した 24 次元のベクトル

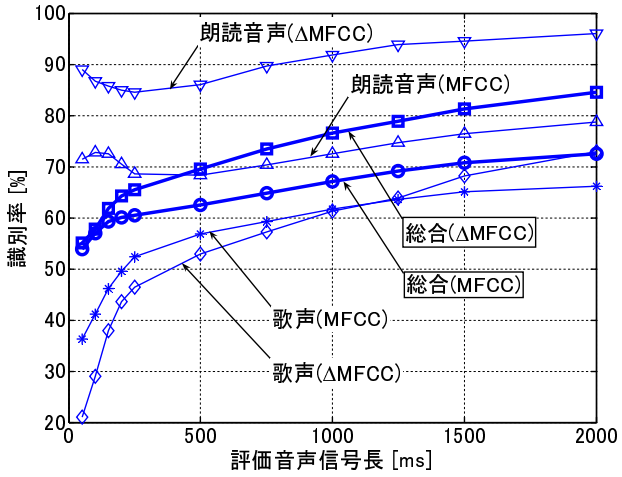

図 6　スペクトル包絡 (MFCC), その時間変化 ΔMFCC を利用した場合の識別率の推移

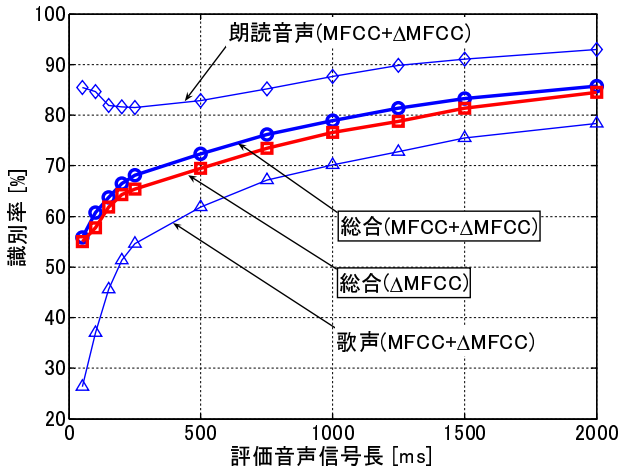

図 7　スペクトル包絡 (MFCC) とその時間変化 ΔMFCC からなる 24 次元ベクトルを利用した場合の識別率の推移

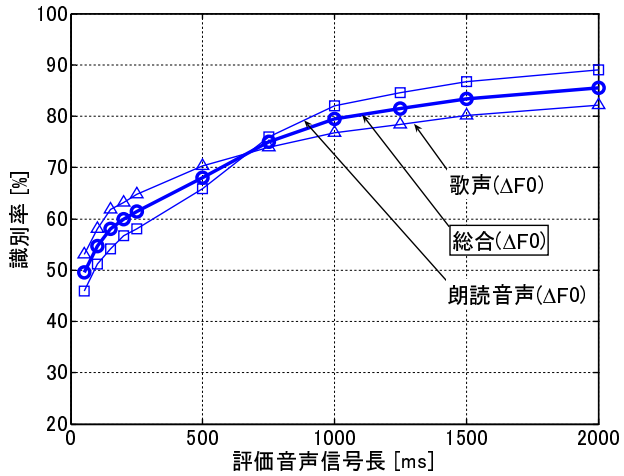

図 8　ΔF0 を利用したときの識別率の推移

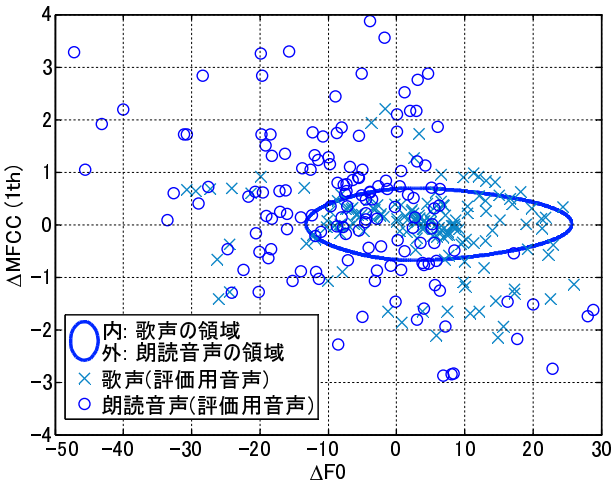

図 9　ΔF0 と ΔMFCC ($1th$) による歌声と朗読音声の識別境界

を用いて歌声と朗読音声の識別を行った場合の識別率の推移である．比較として図 6 に示されている ΔMFCC を利用した場合の結果も合わせて示す．MFCC と ΔMFCC を利用することによって, ΔMFCC のみの場合よりもさらに識別率が上昇していることがわかる．発声開始より 2 秒間の音声信号に対して 84.5%の識別率が得られた．

5.3　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手法に関しても, 表 3 に基づいて同様の学習・評価音声を利用して評価を行った．

図 8 は ΔF0 の GMM による識別結果を表している．発声開始からの音声信号長が長くなるにつれて単調に識別率が上昇していくことがわかる．また, 音声信号長が 700ms 以上で, 朗読音声の識別率が歌声の識別率を上回っていることがわかる．これは, 朗読音声の F0 の下降が, 発声開始より約 700ms 程度から始まることを意味しているのではないかと考えられる．図 4 からわかるように, 朗読音声の F0 の軌跡は発声開始では一度上昇し, への字型に下降が始まる．しかし, 図 3 では, F0 が上昇し, 時間が経過しても同じ F0 を維持している様子がわかる．これは歌声がメロディにより制約を受けているためである．このように短時間の局所的な変化ではなく, 大局的に軌跡を眺めることによって, 歌声と朗読音声の違いを捉えることができたと考えられる．発声開始より 2 秒間の音声信号に対して 85.0%で 2 つの音声の識別が可能である．

5.4　局所的特徴と大局的特徴の比較

図 9 は, ΔF0 と ΔMFCC ($1th$) による歌声と朗読音声の識別境界とともに, 評価用の歌声, 朗読音声, 各 1 サンプルから算出される特徴量 (ΔF0, ΔMFCC ($1th$)) をプロットしたものである．識別境界の内側は歌声, 外側は朗読音声の領域となる．図より歌声は, $-15 < \Delta F0 < 30$ かつ $-1 < \Delta MFCC(1th) < 1$ の値を観測する頻度が高いということがわかる．これは, 歌声の F0 の変化が平均的に 0, もしくは正の傾きをもち, スペクトル包絡の変化が小さいということである．

図 10 は局所的な特徴として MFCC+ΔMFCC, 大局的な特徴として ΔF0 を利用したときの識別結果である．両方の特徴とも, 評価音声の時間長が長くなるにつれて, 識別性能が上昇していくことがわかる．評価音声の時間長が 1 秒よりも短い場合, MFCC, ΔMFCC による識別が有効である．また, ΔF0 は,

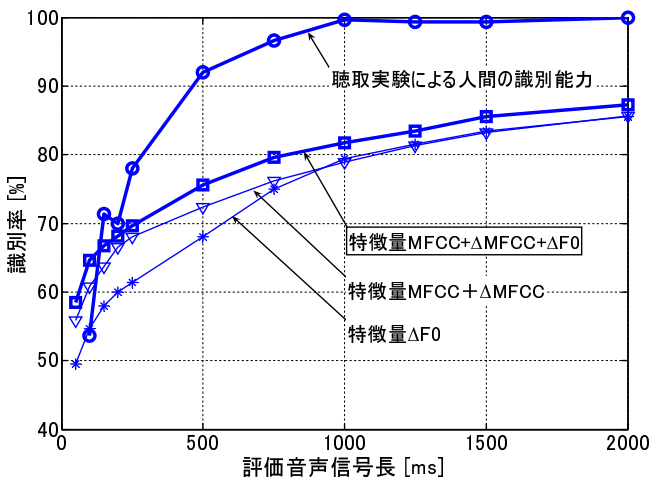

図 10　局所的特徴と大局的特徴による識別率の比較

評価音声信号が 1 秒よりも長い場合に識別に有効であることがわかる．

最終的に 2 つの特徴量を統合した．すなわち MFCC+ΔMFCC+ΔF0 の 25 次元のベクトルによる歌声と朗読音声の識別を試みた．図 10 より 2 つの尺度を統合することによって 2 秒の音声信号に対して識別率は 87.3%となった．

6. 考　　察

実験結果より, 2 つの識別尺度が歌声と朗読音声の識別において, 効果的に音声信号の特徴を捉えていることが明らかとなった．MFCC と ΔMFCC を利用した識別器では, 1 秒以下の音声信号に対して有効である．これは歌声と朗読音声のスペクトル包絡の違いが, 短時間の音声信号の識別に対して優勢な手がかりとなるということである．一方で, ΔF0 を利用した識別器では, 一秒よりも長い音声信号に対して効果的である．これは, ΔF0 が, 歌声と朗読音声の F0 の大局的な軌跡の違いを適切に表現しているからであると考えられる．今回 Δ の時間幅は 50ms と非常に短いが, その局所的な変化分, すなわち歌声の F0 は急激な変化が頻繁に起こらないということ, 朗読音声は, F0 の下降する頻度が高いという特徴を捉えることが, 結局は F0 の大局的な軌跡を見ていることになると考えられる．

7. まとめと今後の課題

本報告では, 歌声と朗読音声に対して, 2 つの異なる側面, すなわち局所的・大局的な特徴をモデル化することによって音声信号の識別を行った．MFCC に基づく提案手法では, 短い音声信号から有効であり, 250ms の音声信号に対して 68.1%の識別率を達成できた．一方で, ΔF0 に基づく提案手法は, 音声信号が 1 秒よりも長い場合に有効であり, 85.0%の識別率が得られた．最終的に, 2 つの尺度を統合することによって, 2 秒の音声信号に対して 87.3%の識別率が得られた．しかし, 聴取実験結果による人間の識別能力と比べたところ, 提案手法の識別性能はまだまだ低く, 人間は 500ms 聞いただけで, 提案手法の 2 秒の場合以上の識別率を持つことがわかった．

今後の課題として, まずは新たな識別特徴量の検討である．今回は周波数領域の特徴量について検討を行ったが, 音声の振幅やパワー, それらの時間変化のような時間領域での音声信号の特徴量については検討を行っていない．これらを用いたとき, どの程度, 識別率が向上するかを検討する必要がある．さらに性能の改善を図る上で, 人間が音声のどのような特徴を手がかりとして識別しているかを知ることも重要である．そこで, 音声信号に様々な変形・破壊を施すことによって, 識別に影響する特徴を聴取実験から検討することも考えている．また, 今回は大量の歌声・朗読音声から学習した GMM に対して, 評価用音声が入力されたときの事後確率の比較によって識別を行っていた．さらにマルコフモデルなどを利用して, 時系列の変化をモデル化する識別手法についても検討する必要がある．最終的には様々な識別特徴量を組み合わせることによる音声の自動識別手法を目標としており, 複数の識別特徴量をどのように統合するかについても検討していきたい．

文　　献

[1] D. Childers and C. Lee. Vocal quality factors: Analysis, synthesis, and perception. *JASA*, Vol. 90, No. 5, pp. 2394–2410, 1991.

[2] W. Chou and L. Gu. Robust singing detection in speech/music discriminator design. In *Proc. ICASSP 2001*, pp. 865–868, May 2001.

[3] Y. Edmund Kim. *Singing voice analysis/Synthesis*. PhD thesis, MIT, September 2003.

[4] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Popular, classical, and jazz music databases. In *Proc. ISMIR 2002*, pp. 287–288, October 2002.

[5] S. Johan. The acoustics of the singing voice. *Scientific American*, p. 82, March 1977.

[6] H. Kawahara and H. Katayose. Scat singing generation using a versatile speech manipulation system, STRAIGHT. *JASA*, Vol. 109, pp. 2425–2426, 2001.

[7] T. Saito, M. Unoki, and M. Akagi. Extraction of F0 dynamic characteristics and development of F0 control model in singing voice. In *Proc. ICAD 2002*, pp. 275–278, July 2002.

[8] T. Saito, M. Unoki, and M. Akagi. Development of the F0 control method for singing-voices synthesis. In *Proc. SP 2004*, pp. 491–494, March 2004.

[9] J. Saunders. Real-time discrimination of broadcast speech/music. In *Proc. ICASSP 1996*, pp. 993–996, May 1996.

[10] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proc. ICASSP 1997*, pp. 1331–1334, April 1997.

[11] C. Shih and G. Kochanski. Prosody control for speaking and singing styles. In *Proc. EUROSPEECH2001*, pp. 669–672, September 2001.

[12] 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, 第 33 巻, pp. 653–658, December 2003.

[13] 後藤真孝, 伊藤克亘, 速水悟. 自然発話中の有声休止箇所のリアルタイム検出システム. 電子情報通信学会論文誌 D-II, 第 J83-D-II 巻, pp. 2330–2340, November 2000.

[14] 後藤真孝, 西村拓一. AIST ハミングデータベース: 歌声研究用音楽データベース. 情報処理学会 音楽情報科学研究会 研究報告, 第 61 巻, August 2005.

[15] 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, 第 34 巻, pp. 41–46, January 2004.