スペクトル包絡と基本周波数の時間変化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[†]

スペクトル包絡と基本周波数の時間変化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 200 ms，1 s 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った．また，識別に影響する音響的特徴について調査するために，短時間のスペクトルの特徴，また韻律を変形させた音声信号を聴取させたところ，それぞれの特徴が相補的に識別の手がかりになることを確認した．この結果より，短時間，長時間の音声信号に対して，それぞれ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）と基本周波数の時間変化の 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 s よりも長い場合，基本周波数の時間変化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 s の音声に対して 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 s よりも短い場合，スペクトル包絡の違いを識別に利用した方が基本周波数の時間変化を利用するよりも識別性能が高い．最終的に，2 つの尺度を単純に統合することによって 2 s の音声に対して 87.3%の識別率を得ることができた．

Discrimination between Singing and Speaking Voices Using a Spectral Envelope and a Fundamental Frequency Derivative

Yasunori Ohishi,[†] Masataka Goto,[††] Katunobu Itou[†,☆] and Kazuya Takeda[†]

In this paper, we discuss the discrimination between singing and speaking voices by using a spectral envelope and a fundamental frequency (F0, perceived as pitch) derivative of voice signals. According to the results of our preliminary subjective experiments, listeners distinguish between singing and speaking voices with the accuracy of 70.0% for 200 ms long signals and 99.7% for 1 second long signals. To examine how humans discriminate between these two voices, we then conducted subjective experiments with singing and speaking voice stimuli whose voice quality and prosody were systematically distorted by using signal processing techniques. The experimental results suggested that spectral and prosodic cues complementarily contributed to the perceptual judgments. By hypothesizing that listeners depend on different cues according to the length of stimuli, we propose an automatic vocal style discriminator that can distinguish between singing and speaking voices by using two measures: a spectral envelope (MFCC) and an F0 derivative. In our experimental results, when voice signals longer than one second are discriminated, the F0-based measure performs better than the MFCC-based measure. On the other hand, when voice signals shorter than one second are discriminated, the MFCC-based measure performs better than the F0-based measure. While the discrimination accuracy with the F0-based measure is 85.0% for two-second signals, simple combination of the two measures improves it by 2.3% for two-second signals.

† 名古屋大学大学院情報科学研究科
Graduate School of Information Science, Nagoya University
†† 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)
☆ 現在，法政大学情報科学部
Presently with Faculty of Computer and Information Sciences, Hosei University

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であるためである．そこ

で，我々は人間の発声する音響事象の自動識別手法を検討している．

本研究では，まずこれらの音響事象の中の歌声と朗読音声の識別に着目する．歌声については多くの研究がなされ，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いえば，歌声には *Singing Formant* と呼ばれる付加的なフォルマントが存在する[1)~3)]．ただこの *Singing Formant* は，オペラ歌手の歌声に観測され，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であるため，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，歌声と日常会話の話し声との識別が可能である．歌声の声質ばかりでなく，歌い方，話し方というスタイルの違いを人間は識別していると考えられる．ただ，従来は歌声か話し声，どちらか一方を対象とした調査が中心で，両者の識別に影響する音響的特徴に関する知見についてはまだ十分に議論されていない．

また近年，さまざまな音楽と音声の識別手法が数多く提案されてきた[4)~6)]．それらの手法を，歌声と話し声の識別に適用することは困難である．なぜなら音楽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．

本研究の目的は歌声と話し声の声質の違いを明らかにすること，また歌い方，話し方という発声のスタイルの違いを明らかにすることである．この研究の応用には，まず，歌声と話し声の 2 つの入力手段を持つ楽曲検索システムが考えられる．従来は，歌声なら歌声単独で検索可能なシステム[7)] が提案されているが，本研究で検討する音声の自動識別手法により，さらに幅広い入力手段による検索システムが実装可能であると考えられる．このほか，音声対話システムにおける発話検出，感情音声の分野への応用も考えられる．

以下，2 章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査する．次に 3 章では，歌声と朗読音声を自動識別するための具体的な手法として，音声のスペクトル包絡，F0 の時間変化の 2 つの特徴を用いた識別手法を提案する．4 章では使用した歌声データベースについて述べ，5 章では評価実験の結果を示す．最後に，6 章で実験結果に対する考察を述べ，7 章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の調査

歌声と朗読音声の識別に必要な音声信号長と識別に影響する音響的特徴の調査を行った．

表 1 識別に必要な音声信号長の調査における聴取サンプルの構成
Table 1 Listening samples based on signal length in investigation of signal length necessary for discrimination.

音声信号長	歌声	朗読音声
100，150，200，250，500，750，1,000 ms	25 サンプル	25 サンプル
1,250 ms	20 サンプル	20 サンプル
1,500，2,000 ms	10 サンプル	10 サンプル
合計	215 サンプル	215 サンプル

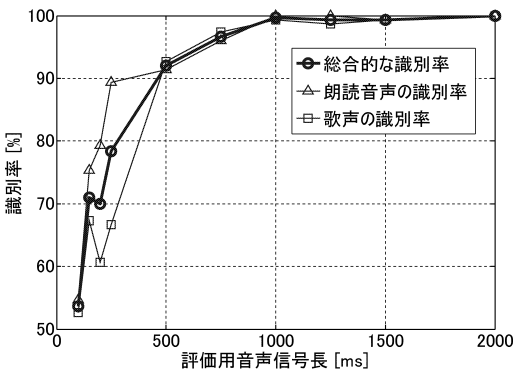

図 1 歌声と朗読音声を人間が聴取して判断する場合の識別率
Fig. 1 Human discrimination performance between singing and speaking voices.

2.1 識別に必要な音声信号長の調査

4 章で述べる歌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，またはその楽曲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 サンプルによる聴取実験を行った．この中から，音声信号 430 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んだ評価セットを 10 種類作る．10 名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，“歌声”，“朗読音声”，もしくは “識別不可能” かの 3 通りで回答させた．**図 1** は聴取した音声信号長に対する識別率の推移である．およそ発声開始から 1 s 程度の音声信号の聴取により，人間は歌声と朗読音声の識別が 100%可能であることが分かる．また，200 ms の長さの音声信号の聴取で，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．さらに人間が音声信号のどのような音響的特徴に着目して識別を行っているかを調査した．

2.2 識別に影響する音響的特徴の調査

1 s の音声信号に対して，人間は 99.7%で識別が可能であることを確認した．そこで識別の手がかりとなる音響的特徴を調査するために，**図 2** のように，1 s

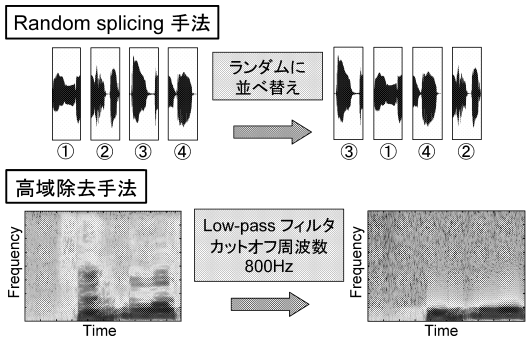

図 2 Random splicing 手法と高域除去手法
Fig. 2 Random splicing and low-pass filtering techniques.

表 2 識別に影響する音響的特徴の調査における聴取サンプルの構成
Table 2 Listening samples in investigation of acoustic cues necessary for discrimination.

Random Splicing 手法		
分割する長さ	歌声	朗読音声
125 ms	40 サンプル	40 サンプル
200 ms	40 サンプル	40 サンプル
250 ms	20 サンプル	20 サンプル
合計	100 サンプル	100 サンプル
高域除去手法		
	歌声	朗読音声
合計	100 サンプル	100 サンプル

の音声信号に含まれる言語，非言語情報を識別に使用できないように信号処理により変形させた 2 種類の音声信号を用いて聴取実験を行う．

Random Splicing 手法[8),9)]

音声区間をある長さの断片に分割し，ランダムに接合することによって，音声の時間的に変化する韻律を変形させる手法である．歌声本来のメロディとリズムのパターン，朗読音声のイントネーション，またそれぞれの発声速度が変形されるが，音声の短時間のスペクトルの特徴は保持された音声信号を聴取したときの識別能力の調査を行う．

高域除去手法

低域通過フィルタにより音声信号の高調波成分を除去し，音質を低下させる．フィルタのカットオフ周波数は 800 Hz とした．すなわち，時間構造は保たれるが，短時間のスペクトルの特徴が変形された音声信号を聴取したときの識別能力の調査を行う．

女性 25 名，男性 25 名の発声開始から 1 s の音声信号に対して 125 ms，200 ms，250 ms の分割長で Random Splicing した音声 15,000 サンプル，高域除去した音声 5,000 サンプルを用意した．これらの 2 つの音声サンプル群の中から，**表 2** のようにそれぞれ 200 サンプルをランダムに選んだ評価セットを 10 種類作

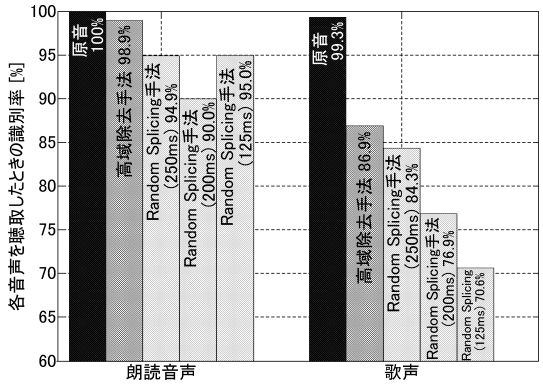

図 3 Random Splicing，高域除去した音声信号の識別率
Fig. 3 Discriminative rate of voice signals by Random Splicing technique and low-pass filtering technique.

表 3 正しく答えられた音声信号の回答に対する確信度
（5 段階評価）
Table 3 Confidence level (1 to 5) for own answer given to each stimulus.

変形手法	歌声	朗読音声
Random Splicing 手法 (250 ms)	4.32	4.06
Random Splicing 手法 (200 ms)	3.87	3.61
Random Splicing 手法 (125 ms)	3.66	3.35
高域除去手法	3.58	4.56

る．10 名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，“歌声”，“朗読音声” の 2 通りで回答させた．また，その回答の確信度を 5〜1 で評価させた．つまり回答に自信があれば 5 を，自信がなければ 1 を評定することになる．

2.3 聴取実験の結果と考察

変形させた音声信号に対する聴取実験結果を**図 3** に示す．

2.3.1 Random Splicing による聴取実験結果

図 3 より原音，すなわちまったく変形を施していない 1 s の朗読音声，歌声の識別率は 100%，99.3%であるのに対して，Random Splicing することによって識別率は低下した．特に歌声は Random Splicing の分割長を短くするにつれて識別率が著しく低下した．分割長 125 ms の場合，歌声の識別率は 70.6%であり，原音の識別率と 28.7%の差が見られる．一方，分割長 125 ms の朗読音声の識別率は 95.0%であり，原音の場合に比べて 5.0%の低下にとどまった．また，**表 3** は正答した聴取サンプルの確信度を示す．分割長を短くするにつれて，その回答に対する確信度，すなわち信頼性が低下することが分かる．実験後の被験者の感想によると，「歌声の伸ばす発声に着目」「音声の音色に着目」「声の大きさの変動が大きければ歌声」「音声

信号内の音高の変動が大きければ歌声」「女性音声の方が朗読音声と歌声の音高差が大きく識別しやすい」という意見を得た．

2.3.2 高域除去手法による聴取実験結果

図 3 より高域除去した朗読音声の識別率は 98.9%，歌声の識別率は 86.9%であった．Random Splicing の場合と同様に歌声の方が朗読音声に比べて識別率の低下が大きい．実験後の被験者の感想によると，「テンポ，発声速度，リズムの違いに着目」「音高が一定の箇所があれば歌声」「音高の変化の違いに着目」という意見を得た．

2.3.3 考　　察

Random Splicing 手法により歌声の識別率が低下した理由は，歌声の持つ本来のメロディのパターンが識別に使用できないように変形されたためであると考えられる．また，分割長が短くなるにつれて，歌声特有の母音を伸ばす発声が細かく切断されてしまい，長さの短い母音の数が増え，朗読音声の母音長と区別がつかず，誤識別してしまったとも考えられる．歌声を朗読音声と誤識別されたある聴取サンプルを分析したところ，原音では母音の長さが平均 146.7 ms であったものが，分割長 125 ms の Random Splicing により母音の長さが平均 73.3 ms と半分の長さになった．一方，同じ歌詞を朗読した音声では，平均 70 ms の母音が，Random Splicing により平均 60 ms となり，それほど母音の長さの変化が見られない．したがって，朗読音声は Random Splicing しても朗読音声と聞こえるが，歌声は Random Splicing することによって朗読音声に聞こえてしまうことが明らかになった．

また，高調波成分を除去したとしても，朗読音声はイントネーションやテンポの違いから識別が可能であると考えられる．しかし，歌声の識別は必ずしも容易ではなく，その短時間のスペクトルの特徴も識別に必要なのではないかと考えられる．

3. 識 別 尺 度

聴取実験結果より，歌声と朗読音声の識別には，母音の長さとその短時間のスペクトルの特徴，また韻律の変化，発声速度が大きく影響することが分かった．また，1 s 程度の音声信号が与えられれば，100%識別可能である一方で，音声信号が 200 ms 程度の短時間であったとしても 70.0%の識別率が得られた．すなわち，短時間，また長時間に観測される特徴が，互いに相補的に影響することによって音声が識別されているということを想定し，本研究では 2 つの異なる尺度を提案する．**図 4** は，被験者がある楽曲を歌い，またその歌詞を朗読した音声のスペクトログラムである．歌声のスペクトログラムでは，広帯域にわたって倍音構造が鮮明に現れ，F0 が朗読音声とは異なる軌跡を描いていることが分かる．また図 4 の右下には，各音声のスペクトログラム中の枠で囲んだ部分の平均スペクトルを示す．この部分は同一の音素/e/を発声しており，歌声は 2〜4 kHz のあたりのパワーが強く，スペクトル包絡の違いが分かる．そこで，音声信号のスペクトル包絡と，F0 の時間変化を利用した識別尺度を提案する．

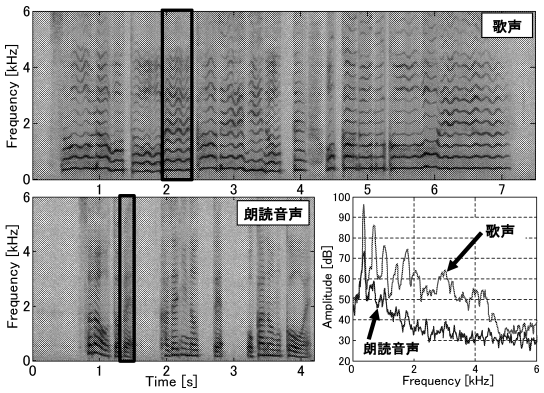

図 4 歌声と朗読音声のスペクトログラム（同一歌詞を発声）
Fig. 4 Spectrogram of singing and speaking voices corresponding to the same lyrics.

表 4 音声の分析条件
Table 4 Analysis conditions of the voice signal.

標本化周波数	16 kHz
分析窓	ハミング窓
フレーム長	25 ms
フレームシフト	10 ms
メルフィルタバンク数	24
使用帯域	0〜8,000 Hz

3.1 スペクトル包絡に基づく尺度

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．音声の分析条件を**表 4** に示す．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．

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

3.2 F0 の時間変化に基づく尺度

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り，楽曲の音符に従った階段構造を形成する．一方で，日

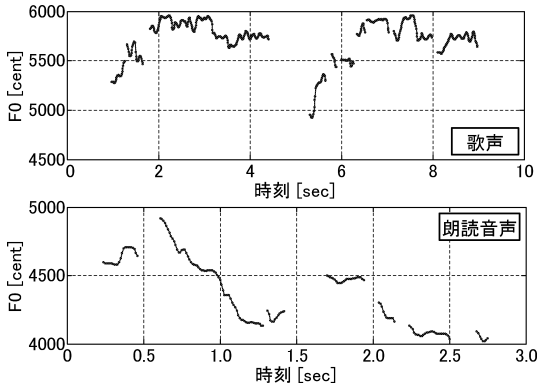

図 5 歌声と朗読音声の F0 の軌跡（同一歌詞を発声）
Fig. 5 F0 contour of singing and speaking voices corresponding to the same lyrics.

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（**図 5**）．それゆえに音声信号から抽出される F0 の軌跡の違いをとらえることは，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，F0 の時間変化 ΔF0 を利用する．

3.2.1 F0 抽 出

F0 は，後藤ら[10] の提案した有声休止検出のための F0 推定手法を利用して，10 ms ごとに推定した．この手法は，非周期的な雑音に加え，高調波構造を持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 F0 を，音声の F0 として抽出する．

3.2.2 ΔF0 の算出

ある時間幅にわたって計算された ΔF0 を歌声と朗読音声の識別に利用する．すなわち，図 5 において，発声区間ごとに F0 を切り出し，各発声区間で式 (1) を利用して，連続した $2K+1$ 個の F0 から回帰係数を計算する．

3.3 識別モデルの学習

歌声，朗読音声それぞれの MFCC，ΔMFCC ベクトル，ΔF0 の分布を 16 混合ガウス分布でモデル化する．ガウス分布の混合数は実験的に決定した．混合ガウス分布の共分散行列は対角共分散行列を利用する．以下のように平均対数事後確率を最大にする音声を識別結果とする．

$$\hat{d} = \underset{d=\text{歌声，朗読音声}}{\operatorname{argmax}} \frac{1}{N} \sum_{n=1}^{N} \log f(\mathbf{x}_n; \Lambda_d) \qquad (2)$$

ここで $\mathbf{x}_n$ は入力音声信号から得られる n 番目の特徴ベクトル，N は特徴ベクトル系列の長さ，Λ_d は歌声，朗読音声の識別特徴量の分布を GMM でモデル化したときのパラメータ（各ガウス分布の重み，平均，共分散行列）である．関数 f は n 番目の特徴ベクトルに対して，歌声，朗読音声それぞれの GMM パラメータを利用したときの事後確率を算出するものとして定義する．

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」[11] の一部である，日本人歌唱者 75 名分（男性 37 名，女性 38 名）の音声データを使用した．各歌唱者が，「RWC 研究用音楽データベース：ポピュラー音楽」（RWC-MDB-P-2001）[12] 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 7,500 サンプルとなる．音声サンプルの長さの平均は歌声で平均 12.0 s，朗読音声で平均 7.0 s であった．

5. 提案手法の評価

本章では，まず音声信号の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 F0 の時間変化，すなわち ΔF0 による識別手法を評価する．最後に 2 つの識別尺度の性能を比較し，それらを組み合わせた手法も評価する．前章で紹介した歌声データベースに含まれる話者を 3 グループ，楽曲を 5 グループに分け，15 回のクロスバリデーションで評価を行った．

5.1 スペクトル包絡を利用した識別性能

スペクトル包絡を利用した歌声と朗読音声の識別性能を評価する．

図 6 は，MFCC と ΔMFCC を利用したときの歌声と朗読音声の識別結果である．MFCC は 12 次までの係数を利用し，ΔMFCC を算出する時間幅は実験的に前後 2 フレームの計 5 フレーム（式 (1) で $K=2$）とした．このとき，総合（平均）識別率が，評価音声信号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分かる．MFCC を利用した場合，2 s の評価音声信号に対して識別率が 72.4%であった．一方，ΔMFCC を利用した場合，2 s の評価音声信号に対して識別率は 84.5%であり，MFCC の識別率を 12.1%上回った．

図 7 は，MFCC と ΔMFCC を連結した 24 次元のベクトル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 に示される ΔMFCC のみを利用した場合の結果もあわせて示す．MFCC と ΔMFCC のベクトルを連結することによっ

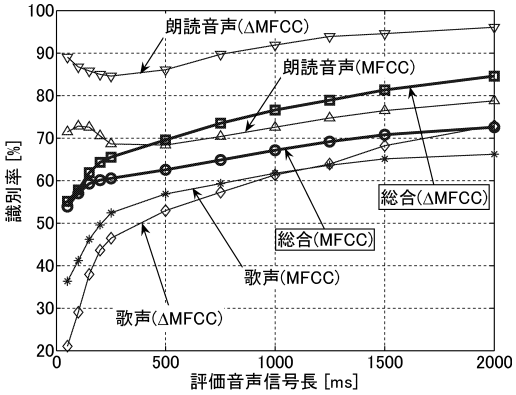

図 6 スペクトル包絡（MFCC），その時間変化 ΔMFCC を利用した場合の識別率の推移
Fig. 6 Discrimination accuracy using short-term features (GMMs of MFCC and ΔMFCC).

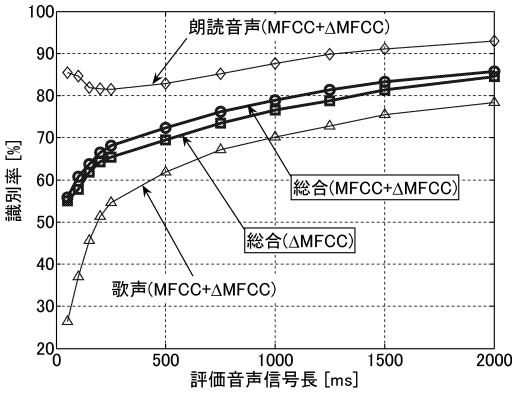

図 7 スペクトル包絡（MFCC）とその時間変化 ΔMFCC からなる 24 次元ベクトルを利用した場合の識別率の推移
Fig. 7 Discrimination accuracy using short-term features (GMMs of MFCC+ΔMFCC 24-dimensional vector).

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分かる．発声開始より 2 s の音声信号に対して 84.7%の識別率が得られた．

5.2 F0 の時間変化を利用した識別性能

F0 の時間変化を利用した識別手法に関しても，同様の学習・評価音声を利用して評価した．ΔF0 を算出する時間幅は実験的に前後 2 フレームの計 5 フレーム（式 (1) で $K = 2$）とした．

図 8 に ΔF0 の GMM による識別結果を示す．発声開始より音声信号長が長くなるにつれて単調に識別率が上昇し，2 s の音声信号に対して 85.0%で識別が可能であることが分かる．

5.3 識別尺度の統合

最後にスペクトル包絡と F0 の時間変化からなる 2 つの識別尺度を統合した識別結果を示す．すなわち

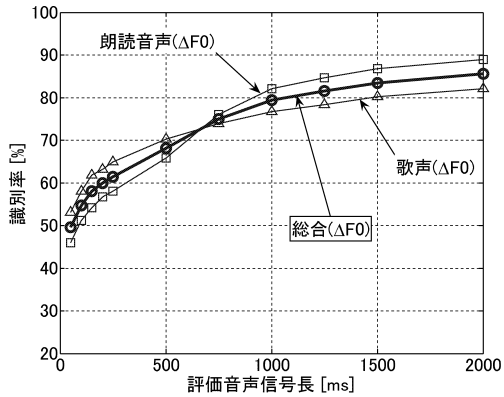

図 8 ΔF0 を利用した場合の識別率の推移
Fig. 8 Discrimination accuracy using ΔF0 GMM.

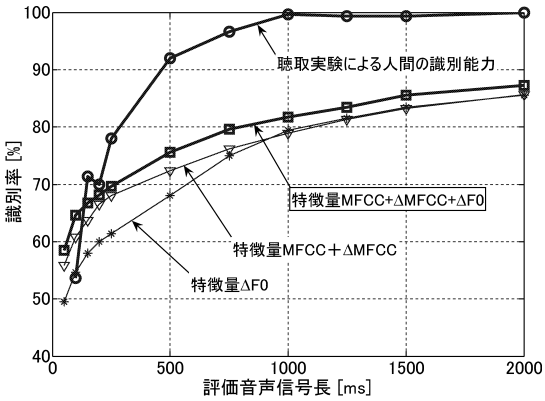

図 9 2 つの識別尺度の比較と統合
Fig. 9 Comparing and integrating two measures using a spectral envelope (MFCC) and ΔF0.

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る．**図 9** より，識別尺度を統合することによって 2 s の音声信号に対して識別率は 87.3%となり，単独の尺度を用いた場合に比べて，最大 2.6%の性能改善が得られた．

6. 考　　察

前章で示したスペクトル包絡と F0 の時間変化を利用した識別結果に対する考察を行う．図 6，7 より，MFCC ベクトルに ΔMFCC を加えることによって識別性能が向上した．これはスペクトル包絡の時間変化にも，音声を識別するための手がかりが含まれているのではないかと考えられる．すなわち，歌声は母音を伸ばす発声からスペクトル包絡の時間変化が小さい．一方で，朗読音声は次々と音素が遷移するのでスペクトル包絡の時間変化が歌声に比べて大きいと考えられる．

図 8 では，歌声，朗読音声それぞれの識別率の値に大差がなく，ともに識別性能が向上していくことが分かる．Δ 算出の時間幅は 50 ms と非常に短いが，GMM から算出されるフレームごとの事後確率を累積して観測することによって，F0 の軌跡の違いをとらえることができたと考えられる．

図 9 より，評価音声の信号長が 1 s よりも短い場合，MFCC，ΔMFCC による識別が有効である．このことから歌声と朗読音声のスペクトル包絡，またその時間変化の違いが，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考えられる．一方で，ΔF0 は，評価音声信号が 1 s よりも長い場合に識別に有効である．これは，ΔF0 によって歌声と朗読音声の長時間に観測される特徴の違いを適切に表現しているからであると考えられる．以上より，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とらえていることが明らかとなった．

最後に 200 ms という非常に短い音声信号の識別と 1 s の比較的長時間の音声信号の識別に着目して，聴取実験結果と自動識別手法の性能との比較を行う．Random Splicing することにより，音声信号の時間構造を変形させたところ，母音長が本来の歌声の母音長より短くなり，韻律的側面の除去のため，人間の歌声の識別能力が低下したと考えられる．しかし，それでも 200 ms の断片化では 70.6%の識別率が得られている．これは，音声信号の短時間のスペクトルの特徴が識別に影響したと考えられる．このことを検証するために，人間の 200 ms の音声信号の識別能力と MFCC+ΔMFCC を利用した 200 ms の音声信号の自動識別結果を比較すると，歌声に関しては人間の識別能力に比べて 9.3%低いものの，他の条件の結果よりも識別率が近い結果が得られている（**図 10** の上図）．このことから短時間のスペクトルの特徴を自動識別に利用することの有効性が示された．ただ，今回使用した MFCC は，音声認識では，音素の構造を適切に表現するために十分な特徴量ではあるが，音声の識別を行うためには，MFCC では表現しきれない部分にも着目する必要があると考えられ，今後，MFCC 以外の特徴量を検討する必要がある．

また，高調波成分を除去することにより音声信号の短時間のスペクトルの特徴を変形させたとしても，人間は，音声信号の時間的に変化する特徴を知覚することによってその識別が約 90%程度可能であった．すなわち長時間の音声信号に含まれる韻律の変化が音声の識別には重要であると考えられる．しかし，ΔF0 を利用した自動識別結果と比較すると，朗読音声，歌声

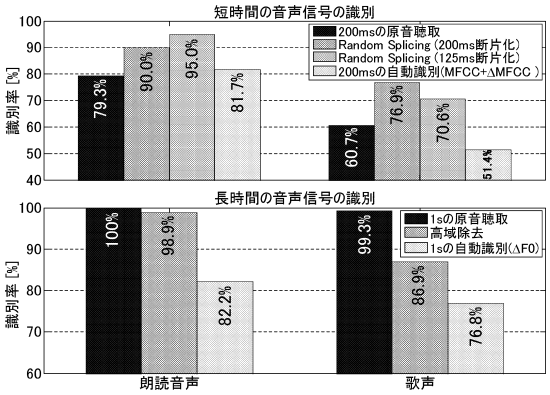

図 10 聴取実験結果と自動識別手法の性能との比較
Fig. 10 Comparison between results of the listening experiment and performances of the automatic discrimination technique.

ともに聴取実験結果に比べて 10%以上低い（図 10 の下図）．これを改善する 1 つの方法としては，さらに長時間の ΔF0 を算出することであると考えられる．今回は，式 (1) を利用して F0 が連続して推定されている部分から ΔF0（Δ の時間幅は 50 ms）を算出した．しかし，聴取実験の結果から，人間はより長時間の F0 の変化を知覚して識別を行っていると考えられる．そこで F0 が連続していない無声音の区間における F0 の補間も考慮に入れたさらに長時間の ΔF0 の算出方法を検討する必要がある．

7. むすび

本研究では，歌声と朗読音声の自動識別手法を提案した．まず，その識別尺度を検討するために聴取実験を行い，人間は 200 ms，1 s 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．また，言語，非言語情報を識別に使用できないように変形させた音声信号を聴取させたところ，音声の短時間のスペクトルの特徴，F0 の時間変化，それぞれが相補的に識別の手がかりになることを確認した．そこで，MFCC を利用して音声のスペクトル包絡を表現し，F0 の時間変化を利用して F0 の軌跡の違いに着目した識別尺度を提案した．MFCC を利用した提案手法では，短い音声信号に対して有効であり，250 ms 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 を利用した提案手法では，音声信号が 1 s よりも長い場合に有効であり，2 s の音声信号に対して 85.0%の識別率が得られた．最終的に，2 つの識別尺度を統合することによって，2 s 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

較したところ，提案手法の識別性能はまだまだ低く，人間は 500 ms 聞いただけで，提案手法の 2 s の場合以上の識別率を持つ．そこで今後は，歌声と朗読音声の音素ごとのスペクトルの特徴の違いを明らかにすること，より長時間の F0 の軌跡を定量的に表現する方法を検討する予定である．

参 考 文 献

1) Kawahara, H. and Katayose, H.: Scat singing generation using a versatile speech manipulation system, STRAIGHT, *J. Acoust. Soc. Amer.*, Vol.109, pp.2425–2426 (2001).
2) Edmund Kim, Y.: Singing voice analysis/Synthesis, PhD Thesis, MIT (2003).
3) Sundberg, J.: Articulatory interpretation of the 'singing formant', *J. Acoust. Soc. Amer.*, Vol.55, pp.838–844 (1974).
4) Scheirer, E. and Slaney, M.: Construction and evaluation of a robust multifeature speech/music discriminator, *Proc. ICASSP 1997*, pp.1331–1334 (1997).
5) Chou, W. and Gu, L.: Robust Singing Detection in Speech/Music Discriminator Design, *Proc. ICASSP 2001*, pp.865–868 (2001).
6) Saunders, J.: Real-time discrimination of broadcast speech/music, *Proc. ICASSP 1996*, pp.993–996 (1996).
7) 園田智也，後藤真孝，村岡洋一：WWW 上での歌声による曲検索システム，電子情報通信学会論文誌，Vol.J81-D-II, No.4, pp.721–731 (1999).
8) Scherer, K.R.: Vocal cues to deception: A comparative channel approach, *Journal of Psycholinguistic Research*, Vol.14, No.4, pp.409–425 (1985).
9) Friend, M. and Farrar, M.J.: A comparison of content-masking procedures for obtaining judgments of discrete affective states, *J. Acoust. Soc. Amer.*, Vol.96, No.3, pp.1283–1290 (1996).
10) 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，Vol.J83-D-II, No.11, pp.2330–2340 (2000).
11) 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，Vol.2005, No.82, pp.7–12 (2005).
12) 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 (2004).

(平成 17 年 10 月 17 日受付)
(平成 18 年 4 月 4 日採録)

大石 康智

2004 年名古屋大学工学部電気電子情報工学科卒業．2006 年同大学大学院情報科学研究科博士前期課程修了．現在，同大学院情報科学研究科博士後期課程．2005 年日本音響学会ポスター賞受賞．日本音響学会会員．

後藤 真孝（正会員）

1993 年早稲田大学理工学部電子通信学科卒業．1998 年同大学大学院理工学研究科博士後期課程修了．同年電子技術総合研究所（2001 年に独立行政法人産業技術総合研究所に改組）に入所し，現在に至る．2000〜2003 年まで科学技術振興事業団さきがけ研究 21「情報と知」領域研究員，2005 年から筑波大学大学院システム情報工学研究科助教授（連携大学院）を兼任．博士（工学）．音楽情報処理，音声言語情報処理等に興味を持つ．1997 年情報処理学会山下記念研究賞（音楽情報科学研究会），1999 年電気関係学会関西支部連合大会奨励賞，2000 年 WISS2000 論文賞・発表賞，2001 年日本音響学会粟屋潔学術奨励賞・ポスター賞，2002 年情報処理学会山下記念研究賞（音声言語情報処理研究会），2002 年日本音楽知覚認知学会研究選奨，2003 年インタラクション 2003 ベストペーパー賞，2005 年情報処理学会論文賞等 18 件受賞．電子情報通信学会，日本音響学会，日本音楽知覚認知学会各会員．

伊藤 克亘（正会員）

博士（工学）．1993 年電子技術総合研究所入所．2003 年名古屋大学大学院情報科学研究科助教授．2006 年法政大学情報科学部教授．現在に至る．音声を主とした自然言語全般に興味を持つ．

武田　一哉（正会員）

1985 年名古屋大学大学院工学研究科修了，同年国際電信電話株式会社（現 KDDI）入社．1986 年国際電気通信基礎技術研究所（ATR）出向．1990 年 KDD 研究所復職（この間 1988〜1989 年まで，米国 MIT 滞在研究員）．1995 年名古屋大学工学部助教授．2003 年名古屋大学情報科学研究科教授．この間，音声コーパス，音声合成，音声認識，音響信号処理，行動信号処理の研究・教育に従事．日本音響学会，電子情報通信学会，IEEE 各会員．