Vol.2012-MUS-94 No.13
Vol.2012-SLP-90 No.13
2012/2/4

スペクトル変化量のピーク間隔・F0・MFCCを用いた歌声と朗読音声の自動識別システム

阿曽慎平[†1] 齋藤毅[†2] 後藤真孝[†3]
糸山克寿[†1] 高橋徹[†1]
尾形哲也[†1] 奥乃博[†1]

本稿では，歌声と朗読音声を識別するシステムについて述べる．入力は無雑音音声，出力は歌声と朗読音声それぞれの尤度（連続値）である．従来，スペクトル包絡（MFCC）と基本周波数（F0）の時間変化に基づいた識別システムが報告されている．これらの特徴量に基づく識別器に，スペクトル変化量のピーク間隔という，音素継続時間に関連する特徴量に基づく識別器を加え，入力音声長に応じて各識別器への重みを変化させた．実験の結果，従来システムでは 1 秒の音声に対し 86.7%の精度であったのに対し，本システムでは 90.2%という結果を得た．本システムが実時間で動作するデモアプリケーションについても述べる．

A System for Automatic Discrimination between Singing and Speaking Voices on the Basis of Peak Interval of Spectral Change, F0, and MFCC

Shimpei Aso,[†1] Takeshi Saitou,[†2] Masataka Goto,[†3]
Katsutoshi Itoyama,[†1] Toru Takahashi,[†1]
Tetsuya Ogata[†1] and Hiroshi G. Okuno[†1]

In this paper we describe a system that discriminates between singing and speaking voices. Given a clean speech signal, it outputs the likelihood of each of the singing and speaking voices. Previous systems use temporal transition of spectral envelope (MFCC) and fundamental frequency (F0) as discrimination features. Our system adds peak interval of spectral change as a phoneme duration feature and weights these features according to the duration of the input speech signal. Experimental results with one-second speech signal show that our system achieves 90.2 % accuracy compared to 86.7 % with previous systems. We also describe a real-time application demonstrating our system.

1. はじめに

人間は話声，歌声，笑い声，泣き声などの様々な種類の音声を発することができる．自然なコミュニケーションのためには，音声の文字情報を理解するだけでなく，音声の種類を聞き分けることが望ましい．人はそれらを聞き分けて適切な応答をすることで，コミュニケーションを成立させているが，計算機システムも同様のことができれば，様々な可能性が広がる．例えば，歌声と話声（朗読音声を含む）を識別して異なる反応をとることができるような未来の音声エンターテイメントシステムとして，音楽再生機能付きカーナビゲーションシステムを考えてみる．ユーザが普通に「渋滞情報を教えて」と話しかけるとシステムは要求に応え，ユーザの気分が高まって歌を歌うとシステムはその楽曲を検索してカラオケ再生する，というような機能が，歌声と話声の自動識別システムがあれば実現できる可能性がある．

本研究では，歌声と話声（朗読音声）の自動識別が将来実用化されることを目指して，その精度向上に取り組む．現在の音声を扱えるシステムの殆どは，話声（朗読音声）又は歌声のいずれか一方が入力されることを前提としている．朗読音声を用いたナビゲーション機能と歌声を用いた楽曲検索機能[1] の両方を備えたシステムであっても，他方の機能を使う際にはスイッチで操作したり，特定の発話をしたりする必要があり，入力モードを意識しないシームレスな使用は困難であった．

従来，大石ら[2] は歌声と朗読音声の識別に必要な特徴量を調査する聴取実験を行い，人間は長時間・短時間の両方の特徴を手がかりとしていることを指摘した．また，声質や音高，及びそれらの時間変化が識別に影響を与えることを想定して，メル周波数ケプストラム係数（Mel-Frequency Cepstram Coefficients，MFCC）と基本周波数（F0）の時間変化を利用した歌声と朗読音声の識別システムを実装し，2 秒の音声に対し 87.3%の精度を得ていた．

しかし，大石らのシステム[2] では 100 ミリ秒以下の比較的短時間の手がかり（音響特徴量）のみを利用しており，入力音声の長さにかかわらず各特徴量の重要度（重み）は一定で

†1 京都大学 大学院情報学研究科
Graduate School of Informatics, Kyoto University
†2 金沢大学 理工学域 電子情報学類
School of Electrical and Computer Engineering, College of Science and Engineering, Kanazawa University
†3 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

Vol.2012-MUS-94 No.13
Vol.2012-SLP-90 No.13
2012/2/4

あった．人間は音素継続時間のような長時間の特徴量も利用しており[2),3)]，音声の長さに応じて手がかり（音響特徴量）を使い分けている可能性が考えられる．そこで本研究では，短時間・長時間の音響特徴を併用し，入力音声の長さに応じて各特徴量の重みを適切に設定することで識別精度を向上できると考えた．

本稿では，歌声と朗読音声の識別器を高精度化する手法と，それを基に開発した実時間で動作するアプリケーションについて述べる．高精度化のアイデアは次の 2 つからなる．

(1) 識別特徴量にスペクトル変化量のピーク間隔を加える．これは数秒の音響信号から抽出される長時間の特徴量で，音素継続時間と必ずしも同一ではないが関連した値を持つ．スペクトル変化量の抽出には Klapuri ら[4)] が提案したアクセント（phenomenal accent）を利用する．

(2) 入力音声長に応じて識別器の重みを可変にする．識別特徴量毎に歌声・朗読音声それぞれの尤度を出力する識別器を構成し，各識別器の出力の重み付け和をとることで最終的な尤度とする．その最適な重みは入力音声長によって異なるので，学習データから事前に求めて設定する．

以下，2 節では関連研究として従来法による識別手法や，精度向上の手がかりとなる既存研究について述べる．3 節で，我々が扱う問題と，解決のためのアイデアを示し，4 節で，それを基に開発したシステムについて述べる．5 節で，我々の手法と従来手法での自動識別精度の比較実験を行う．6 節で実験結果に対する考察について述べ，7 節で，我々の手法を用いた実時間で動作するアプリケーションを紹介する．最後に 8 節で本稿のまとめを行う．

2. 関 連 研 究

本節では，従来手法による歌声と朗読音声の識別手法の概要と問題点について述べる．また，解決のアイデアとなる文献を紹介する．

2.1 MFCC・ΔMFCC・F0 を利用した識別

大石らは MFCC（12 次元）・ΔMFCC（12 次元）・F0（1 次元）を利用した歌声と朗読音声の自動識別手法を提案した[2)]．この識別器の設計にあたり，100 ミリ秒から 2000 ミリ秒の範囲で切り出した歌声・朗読音声や，Random Splicing[5),6)] と呼ばれる手法で音声を 125 ミリ秒から 250 ミリ秒の範囲で細切れにし，ランダムに接合した音声を刺激音を用いて聴取実験を行い，人間の歌声と朗読音声の識別能力に関して以下の報告をしている．

(1) 1 秒の音声を聴いて 99.7%の精度で歌声と朗読音声を識別できる（長いほど識別が容易）

(2) 母音の長さ・母音の短時間のスペクトル特徴・韻律の時間変化が手がかりである

(3) 短時間・長時間特徴の両方が影響を与える

(1) から，人間が高い識別能力を持つことわかる．人間の識別特性を計算機上で実現できれば，高精度な自動識別が可能になると考えられる．(2) の結果から MFCC・ΔMFCC・ΔF0 を識別特徴量としている大石らが提案した学習・識別方法[2)] を以下に述べる．

2.1.1 学　　習

学習時には，特徴量の分布を 16 混合ガウス分布（GMM）でモデル化し，歌声・朗読音声それぞれでパラメータを学習することで構成する（**図 1**，学習時を参照）．共分散行列は対角成分のみを利用する．

2.1.2 識　　別

識別（テスト）時には，入力音声から特徴ベクトルを抽出し，2.1.1 節の方法で学習した識別器にかけ，以下のように平均対数尤度（以後単純に尤度と略記する）が大きい方を識別結果 $\hat{d}$ とする（図 1，識別時を参照）．

$$g_{ad}(\boldsymbol{x}) = \frac{1}{N}\sum_{n}^{N} \log f\left(\boldsymbol{x}_a(n); \Lambda_{ad}\right) \tag{1}$$

$$\hat{d} = \underset{d=\text{歌声, 朗読音声}}{\operatorname{argmax}} g_{ad} \tag{2}$$

$\boldsymbol{x}$ は入力音声，$\boldsymbol{x}_a(n)$ は，入力音声 $\boldsymbol{x}$ から抽出した特徴量 a における先頭から n 番目（つまりフレームインデックスが n）のベクトル，N は入力音声から得られる特徴ベクトルの時系列の個数．Λ_{ad} は，d ($\in$ 歌声，朗読音声) の特徴量 a 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 は入力 $\boldsymbol{x}$ に対し GMM パラメータ Λ を利用した時の尤度を算出する．

2.1.3 結果と課題

MFCC（12 次元）・ΔMFCC（12 次元）・ΔF0（1 次元）を結合した 25 次元のベクトルを特徴量として学習し，識別を行ったところ，2 秒の音声に対して 87.3%の精度が得られることを報告している[2)]．

しかし，この識別手法には次の 2 つの問題がある．

(1) 瞬時的（短時間）な特徴（100 ミリ秒以下）の分布を扱っているため，長時間に渡る特徴を扱えない．人間は短時間だけでなく，長時間の特徴量も利用していると考えられ，それにより高精度な識別を実現していると考えられる．

Vol.2012-MUS-94 No.13
Vol.2012-SLP-90 No.13
2012/2/4

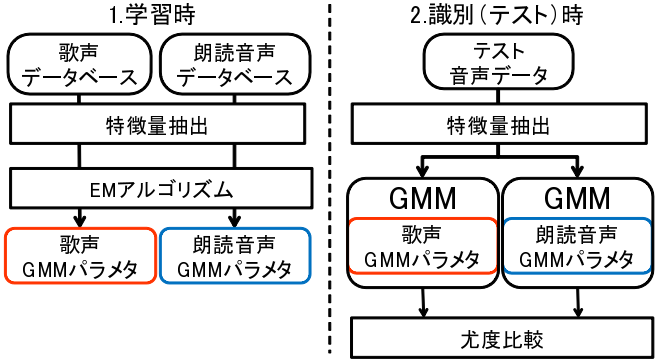

図 1 識別器の学習・テスト方法．学習時に得たパラメータを基に識別器を構成，出力尤度の平均値が大きくなるほうを識別結果とする．

（2） 各特徴量に対する信頼度は常に一定であった．入力音声長によっては，いずれかの特徴量が悪影響を与える場合がある．特徴量毎に識別器を構成し，入力音声長に応じて各識別器の信頼度（重み）を適切に設定すれば，精度の向上が見込める．

これらの課題に対する我々の解決法を 3 節で述べる．

また，大石らは F0・音素継続時間を変化させる際に，音声を細切れにし，並び替え（Random Splicing）を行った．この手法では F0・音素継続時間以外の特徴量にも変化を与えてしまうため，想定外の特徴量の変化し，実験結果に影響を与えていた可能性がある．音声制御ツールを用いて F0・音素継続時間を変化させて実験を行えば，想定外の特徴量に与える影響が小さくなり，より正確な知見が得られると考えられる．2.2 節では，我々が行った聴取実験について述べ，大石らの主張の正当性を確認する．

2.2 歌声らしさ・話声らしさの変化の調査

我々は先行研究で聴取実験を行い，人間は F0 と音素継続時間の両方を識別の手がかりとしていることを確認した[3]．また，前節 2.1 の結果とつき合わせることで，人間は短時間・長時間の音声に対して，異なる特徴量を手がかりとする可能性が指摘できる．この聴取実験では，2.1 節で提案された F0・音素継続時間や，その他の歌声研究で着目されてきた特徴量を置換制御した際に，人間の歌声・朗読音声の識別結果がどう変化するかを調査し，それぞれの特徴量の影響度の順位付けを目的としていた．

F0・音素継続時間・声の Jitter・スペクトル包絡（声質に関連する特徴量）・パワーという 5 つの特徴量を，歌声と朗読音声の間でそれぞれ独立に置換して 72 種類の音声を合成し，聴取実験を行って，それぞれの音声が歌声に聞こえるか話声に聞こえるか調査した．ここで，声の Jitter とは．F0 のゆらぎ成分である[7]．特徴量の置換には音声分析合成ツール STRAIGHT を用いた[8]．また，合成された音声は，大石らの聴取実験（0.1 秒から 2 秒の 10 段階）に比べ長時間であった（5 秒と 1 秒の 2 段階）．

実験結果から，人間は長時間の音声に対して，F0 と音素継続時間を手がかりとして識別していることが確認された．2.1 節から，人間は 1 秒以上の音声に対して歌声と朗読音声を正しく識別できることがわかる．歌声の F0 と音素継続時間のみを朗読音声のものに置換した合成音に対し，ほぼ全ての被験者が朗読音声である回答した．逆に朗読音声の声質を保ったまま F0 と音素継続時間を歌声のものに置換した合成音に対しては，ほぼ全ての被験者が歌声であると回答した．その他の特徴量を置換した場合にはこのような結果は得られなかった．F0・音素継続時間が識別に決定的であり，その他の特徴量は相対的に重要ではなかった考えられる．この結果は，歌声においてメロディやリズムが重要であるという直感的な解釈と合致し，これが定量的に確認されたと考えられる．

また，人間は短時間・長時間の音声に対して，異なる特徴量を手がかりとする可能性が指摘できる．2.1 節の実験では，短時間のスペクトルの特徴が識別に与えていたという結果が得られたが，本節の実験でスペクトル包絡（短時間のスペクトル特徴に関連する特徴量）が識別にほとんど影響を与えなかった．実験に用いた音声が比較的長かったため，F0 や音素継続時間を重視していた可能性がある．

2.3 スペクトル変化量（アクセント）の算出手法

本稿では，スペクトル変化量を，Klapuri ら[4] が提案したアクセント（phenomenal accent）として算出する．このアクセントは，実時間での算出が可能であり，歌声からの自動採譜にも利用されている[9]．そして，算出結果のスペクトル変化量のピーク間隔，つまり，アクセントピーク間隔を識別に用いる特徴量（以後，ピーク間隔と略す）とする．このピーク間隔は，音響的な変化が顕著な箇所を境界として区切った区間の長さであり，音素継続時間と必ずしも同一ではないが関連した値を持つ．例えば，ピーク間隔が短ければ，音素継続時間も短くなる．

Klapuri らのアクセント抽出法[4] は，以下のステップからなる，

（1） 入力音声（標本化周波数 44.1kHz）の平均振幅が 0，分散が 1 となるように正規化した後に離散フーリエ変換をとる（フレーム長 23 ミリ秒，シフト 12 ミリ秒）．

（2） 各フレームに対し 36 個の三角窓の帯域通過フィルターを 50Hz から 20kHz の間に等間隔配置し，各帯域のパワーを算出する．

Vol.2012-MUS-94 No.13
Vol.2012-SLP-90 No.13
2012/2/4

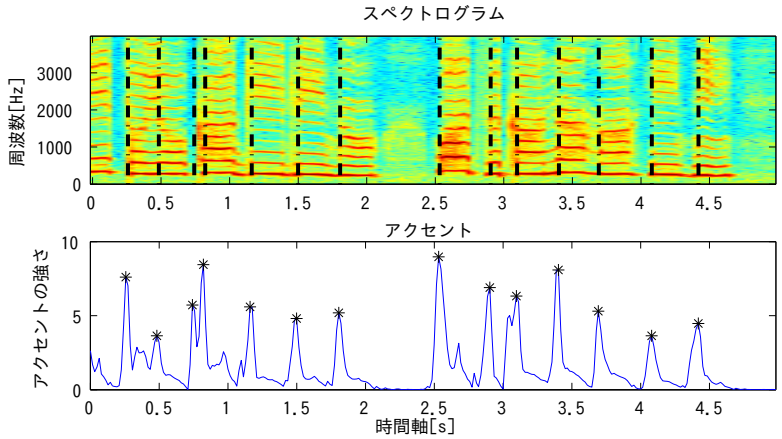

図 2 入力音声のスペクトログラムを表示したもの（上図）とアクセント抽出結果（下図）．下図*印は 4.1 節で述べる方法で抽出したピーク，上図破線は，ピーク位置に合わせて引いた補助線．発声開始や音素が変わる時刻でピークとなることがわかる．

（3） 人間の音圧変化に対する知覚特性を近似するために，各帯域のパワーに対して μ-law compression と呼ばれる圧縮をかける，

（4） 時間解像度をあげるために，補完を行いサンプリング周波数を倍にした $z_b(n)$（b は帯域インデックス，n はフレームインデックス）を得る．

（5） 時間増分量 $z'_b(n)$ を $z'_b(n) = \max(z_b(n) - z_b(n-1), 0)$ の算出結果から得る．

（6） $z_b(n)$ と $z'_b(n)$ の重み付け和を算出する．

（7） 帯域毎の総和をとり，アクセントの強さを得る[9]．

歌声に対して，上記のアクセント抽出法を適用した例を**図 2** に示す．この図から，この特徴量が発声開始時や，音素が変わる瞬間に極大値（ピーク）をもつことがわかる．

3. 本研究のアイデア

我々は新たな音響特徴量を追加し，特徴量の統合方法を改良することで，自動識別の精度向上を目指す．本節では，従来研究での課題を整理し，それらに対する解決方法を述べる．

3.1 課　　題

2 節より，従来手法に関して以下の問題点が指摘できる．

（1） 音素継続時間を用いた歌声と朗読音声の識別手法については，従来あまり議論されてこなかった．人間は音素継続時間を識別の手がかりとすることが報告されており，従って自動識別にも有効だと考えられる．

（2） 入力音声長に応じた特徴量の重み付け処理は行われなかった．人間は短時間・長時間の音声に対して，異なる特徴量を手がかりとする可能性がある．人間のように高い精度で識別するには長時間・短時間の音声に対して異なる特徴量を手がかりにできるような手法が必要だと考えられる．

3.2 アプローチ

これらの問題を，我々は以下の手法で扱う．

（1） 音素継続時間に関連する特徴量を用いる．これには，音素継続時間と関連があり，実時間で抽出可能な，スペクトル変化量（アクセント）のピーク間隔を利用する．

（2） 入力音声長に応じて識別器の重みを可変にする．識別特徴量毎に歌声・朗読音声それぞれの尤度を出力する識別器を構成し，各識別器の出力の重み付け和をとることで最終的な尤度とする．その最適な重みは入力音声長によって異なるので，学習データから事前に求めて設定する．

4. 開発した自動識別システム

本自動識別システムは，従来用いられていた ΔF0・MFCC・ΔMFCC に加え，スペクトル変化量（アクセント）のピーク間隔も特徴量として利用する．いずれの特徴量も実時間で動作可能な方法で抽出する．特徴量毎に歌声尤度から朗読音声の尤度を減算した結果（以後，尤度差と表記する）を出力する識別器を用意し，各識別機の出力の重み付け和をとることで最終的な出力（尤度差）とする．歌声か朗読音声かという 2 値での識別結果が必要な場合は，最終出力の符号（正負）で判定する．

4.1 アクセントピーク間隔の抽出

アクセントピーク間隔は，2.3 節の方法でアクセントを抽出した後，ピークピッキング法を適用して得られた n 番目のピーク時刻 $t(n)$ に対し，隣り合うピーク間隔 $t(n) - t(n-1)$ を算出することで求める．ピークの抽出例を図 2 に示す．1 つのピーク周辺で多数のピークが誤検出されないよう，隣り合うピーク同士の間隔（時間）が T 以上となるように制約 $t(n) - t(n-1) > T$ をかけた．また，小さなピークは無視できるように，発声区間中（アクセントの強さが一定値以上，実験的に決定）の平均アクセント値の P 倍以上になるもののみをピークとして抽出した．T, P の値は実験的に決定（$T = 50$[ミリ秒]，$P = 2.5$）した．

注意点として，アクセントピーク間隔は他の特徴量のようにフレーム（時間）ごとに値が

Vol.2012-MUS-94 No.13
Vol.2012-SLP-90 No.13
2012/2/4

求まるのではなく，アクセントのピークが検出される度にその間隔の値が求まる．そのため，他の特徴量がフレーム数分だけ得られるのに対し，ピーク間隔の特徴量が得られる個数は少なくなる．いずれにせよ，個数にかかわらず，得られた特徴量の分布を学習して識別器を構成する．

4.2 単独特徴量に基づく識別器

2.1 節とほぼ同様の手法を用いて，特徴量の分布から**図 3** のような識別器を構築する．違いは出力形式で，我々は歌声尤度から朗読音声尤度を減算した尤度差（連続値）としている（2.1 節では歌声か朗読音声の識別結果（2 値）であった）．以下の式で表現できる．

$$G_a(\boldsymbol{x}) = (g_{a\ \text{歌声}}(\boldsymbol{x}) - g_{a\ \text{朗読音声}}(\boldsymbol{x})) \tag{3}$$

ここで $\boldsymbol{x}$ は入力音声，a は識別に用いる特徴量である．関数 g_{ad}（$d \in$ 歌声，朗読音声）は 4.2 節で述べた識別器で，音声を入力すると歌声・朗読音声それぞれの尤度を出力するものとして定義する．

本稿では，この識別器を特徴量毎に構築する．4 つの特徴量（アクセントピーク間隔，ΔF0，MFCC，ΔMFCC）を基に尤度差を出力する 4 つの識別器が得られる．4.3 節で述べる識別法と区別するため，これら 4 つの識別器を単独特徴量識別器と呼称する．

4.3 識別器の重み付け統合法

図 4 のように，各単独特徴量識別器から出力される尤度差に対して重み付けをした和を，統合尤度差とする．この統合尤度差 h 及び最終的な識別結果 $\hat{d}$ は以下の式で表現できる．

$$h(\boldsymbol{x}) = \sum_{a \in F} w_a \cdot G_a(\boldsymbol{x}) \tag{4}$$

$$\hat{d} = \begin{cases} \text{歌声} & h(\boldsymbol{x}_a) > 0 \\ \text{朗読音声} & h(\boldsymbol{x}_a) \leq 0 \end{cases} \tag{5}$$

ここで，集合 F は利用する特徴量（本稿ではアクセントピーク間隔・ΔF0・MFCC・ΔMFCC の 4 つ），w_a は各識別器の重み，$\boldsymbol{x}$ は入力音声，$G_a(\boldsymbol{x})$ は 4.2 で述べた単独特徴量の出力（尤度差）である．

重みは学習データから事前に決定する．すなわち，識別器の学習に用いるデータに対する識別精度が高くなるように選ぶ．各識別器に対する重みを 0 から 10 までの 11 段階で変えながら（11^M 通り，M は単独特徴量識別器の数），最も識別精度（2 値判定したときの正解精度）が高くなる組み合わせを選び，総和が 1 となるよう正規化して使用する．

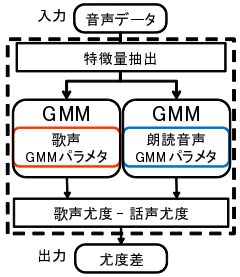

図 3 単独特徴量識別器の構成．歌声と朗読音声の尤度差を出力する．

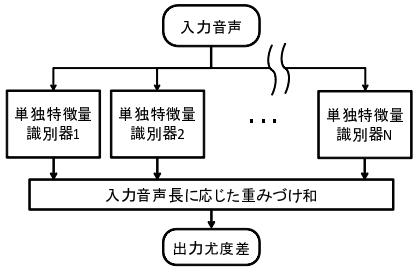

図 4 本統合手法．各単独特徴量識別器の出力に入力音声長に応じたを重み付けた和をとる．

表 1 *F0* の抽出条件

標本化周波数	16kHz
フレームシフト	10 ミリ秒
推定周波数上限	1kHz
推定周波数下限	80Hz

表 2 *MFCC* の抽出条件

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

4.4 その他識別特徴量の抽出

従来手法で利用された識別特徴量の抽出方法について簡潔に述べる．F0 の抽出には yegnanarayana の手法を用いて推定した[10]．この手法は実時間での動作が可能である．F0 の抽出条件を**表 1**，MFCC の抽出条件を**表 2** に示す．F0，MFCC の Δ 成分を算出するために，$2K+1$ 個のフレームにわたる回帰係数を計算した．

$$\Delta c[n] = \frac{\sum_{k=-K}^{K} k \cdot c[n+k]}{\sum_{-K}^{K} k^2} \tag{6}$$

$c[n]$ は Δ 成分を求めたい特徴の n フレーム目の特徴ベクトルである．K の値は 2 とした．

5. 評 価 実 験

アクセントピーク間隔と重み付け統合法による，歌声と朗読音声の識別性能を従来法との比較によって評価する．手法毎に，5.1 のデータベースを利用してクロスバリデーションにより学習・識別を行い，識別結果（歌声または朗読音声の 2 値）と正解ラベルを比較して識

Vol.2012-MUS-94 No.13
Vol.2012-SLP-90 No.13
2012/2/4

別精度を算出する．尚，単独特徴量識別器の識別結果は，尤度の符号（正なら歌声，0 以下なら朗読音声）により決定する．識別性能を評価する手法の一覧を**表 5.3** に示す．

5.1 評価用音声データベース

評価実験には研究用音楽データベース「AIST ハミングデータベース」[11] 中の，日本人による歌声（3750 音）と歌詞朗読音声（3750 音）を使用する．内訳は男性が 37 名，女性が 38 名で，「RWC 研究用音楽データベース: ポピュラー音楽」25 曲の出だしとサビの部分を歌った音声と歌詞を朗読した音声である．音声の平均長は，歌声が 12 秒，朗読音声が 7 秒である．

これらの音声を，話者を 3 グループ，楽曲を 5 グループに分けた 15 回のクロスバリデーションで評価を行う．また，各音声を発声開始時刻から一定時刻切り出して利用する．切り出す時間長を変えながら，識別精度の違いを調査する．

5.2 単独特徴量に基づく識別の精度

アクセントピーク間隔，ΔF0，MFCC，ΔMFCC それぞれを単独で利用した時の歌声と朗読音声の識別精度を比較する．**図 5** に精度と音声長の関係を示す．図から，いずれの特徴量でも，入力音声長に対しほぼ単調増加となっていることが確認できる．また，概ねアクセントピーク間隔，MFCC，ΔF0，ΔMFCC，の順に精度が高い．1 秒・2 秒の音声に対する識別精度はそれぞれ，アクセントピーク間隔が 70.4%・74.8%，MFCC が 71.9%・78.3%，ΔF0 が 78.6%・84.5%，ΔMFCC が 86.3%・91.8%であった．

5.3 統合法の識別精度

以下の，本統合法と従来法（大石らの手法に相当するように独自に実装した手法）の識別精度を比較する．

(1) 大石手法：ΔF0，MFCC，ΔMFCC の 3 つを結合した特徴量を基に識別器を構築
(2) ピーク間隔無し本統合法：ΔF0，MFCC，ΔMFCC の 3 つの単独特徴量識別器を統合
(3) 本統合法：アクセントピーク間隔，ΔF0，MFCC，ΔMFCC の 4 つの単独特徴量識別器を統合

図 7 に，各統合法の精度と音声長の関係示す．また，単独特徴量に基づいた場合との比較図を**図 6** に示す．本統合法で利用した重み係数と音声長の関係を**図 8** に示す，1 秒・2 秒の音声に対する識別精度はそれぞれ，大石らの手法が 86.7%・89.8%，ピーク間隔無し本統合法が 89.5%・94.1%，本統合法が 90.2%・94.6%であった．

表 3 手法一覧．呼称は実験結果図に使用する．

呼称	従来法か本手法か	識別方法説明
ピーク間隔	本手法	アクセントピーク間隔に基づく単独特徴量識別器
ΔF0	従来法	ΔF0 に基づく単独特徴量識別器
MFCC	従来法	MFCC に基づく単独特徴量識別器
ΔMFCC	従来法	ΔMFCC に基づく単独特徴量識別器
本統合法	本手法	上記 4 つの識別器の重み付け和を用いる
ピーク間隔無し本統合法	本手法	ΔF0，MFCC，ΔMFCC の識別器の重み付け和を用いる
大石手法	従来法	ΔF0，MFCC，ΔMFCC を結合した特徴量を用いる

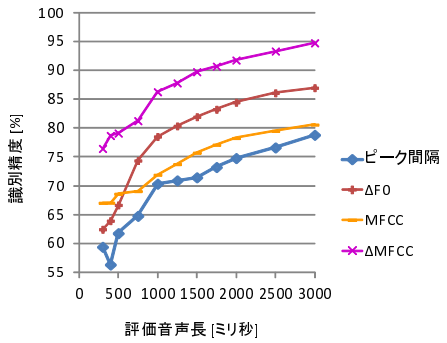

図 5 アクセントピーク間隔，ΔF0，MFCC，ΔMFCC を利用した場合の識別精度と音声長の関係

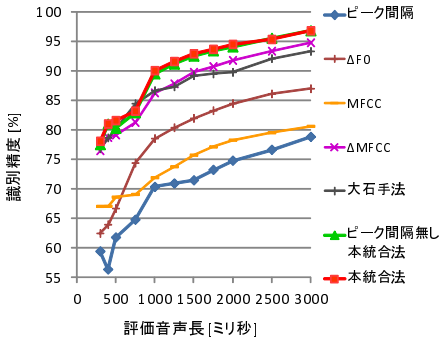

図 6 全ての手法の識別精度と音声長の関係

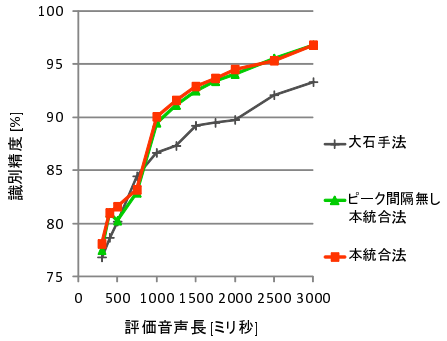

図 7 従来手法・本手法による統合方法と音声長の関係

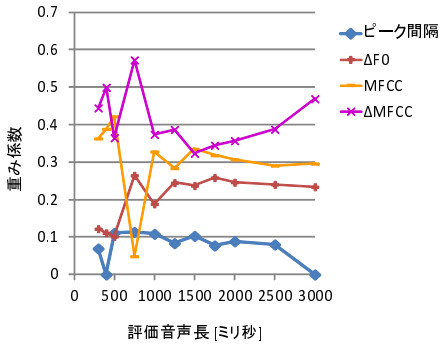

図 8 本統合法に用いた重み係数と入力音声長の関係

Vol.2012-MUS-94 No.13
Vol.2012-SLP-90 No.13
2012/2/4

6. 考　　察

前節の結果から，今回提案したアクセントピーク間隔を用いた識別方法と，重み付け和による統合方法について考察する．図 5 より，アクセントピーク間隔のみを用いた識別手法は従来手法に比べ精度が低いことが判明した．本統合手法は従来手法より高い精度を持つことが示された．図 6，7 を見ると，750 ミリ秒を除く全ての音声長において，本統合法による識別精度が最も高い．図 8 を見ると，この範囲では重みがなめらかに変化しており，各特徴量に対して，適切な重みが設定できたためだと考えられる．その一方で 750 ミリ秒の結果に着目すると，本統合法は従来手法の精度を下回っており，MFCC に対する重みが低くなっている．最適な重みの設定に失敗しているためだと考えられる．

図 8 の重み係数と音声長の関係に着目すると，次のことがわかる．

(1) 1500 ミリ秒から 3000 ミリ秒の範囲では，ΔMFCC に対する重みが増加する傾向が見られる．これは，ΔMFCC が単独の特徴量としては最も高い識別精度をもち，かつ音声長が長くなるにつれ，単独でも十分な精度が得られるためだと考えられる．

(2) 最適な重みは，精度の高い順とは限らない可能性がある．これは，ΔF0 と MFCC を比べると，全体的に MFCC のほうが重みが大きいが，ΔF0 のほうが単独特徴量としては精度が高くなるという結果に反するという点から予想される．

7. デモアプリケーション

本統合法を用いて，実時間で動作するアプリケーションを開発した．このアプリケーションは，我々の開発した識別システムに，音声切り出しモジュールと識別結果視覚化モジュール（表示部）を組み込むことで構成される．マイク（低ノイズを想定しているため，接話タイプが望ましい）を用いて音声入力すると，それが歌声であるか朗読音声であるかを本統合法を用いて判別し，結果（尤度差）をアナログメータ風に表示する．また，アクセントピーク間隔，ΔF0，MFCC，ΔMFCC それぞれ単独で用いた場合の結果（尤度差）も表示する．アプリケーションの概要を**図 9** に示す．以下の構成要素からなる．

(1) Voice Activity Detection（VAD）による音声切り出しモジュール
(2) 各特徴量に基づいた単独特徴量識別モジュール
(3) 入力音声長に応じた重み付け和をとるモジュール
(4) (2)，(3) の尤度差を基に画面に結果を表示するモジュール（図 9 参照）

(1) で用いる VAD は，ゼロ交差数と振幅を元に構成したものであり，発声開始を検知して

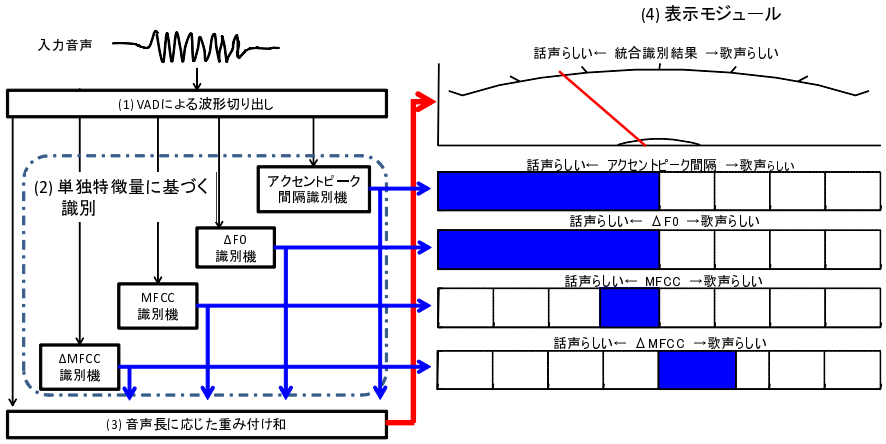

図 9 左：デモアプリケーション概要図．右：表示例．一番上が最終的な識別尤度差（連続値）を示し，下の 4 つが特徴量毎の識別結果（尤度差）を示す．中央より左にいくほど話声尤度のほうが高いことを示し，右にいくほど歌声尤度のほうが高いことを示す．この表示例は話声を入力したもので，ΔMFCC が歌声だと誤識別（歌声のほうが尤度が高い）しているが，他の特徴量では話声の尤度が高いため，統合結果も話声の尤度が高くなることを示している．

録音を開始し，発声が終了するか，発声開始から一定時間（200 ミリ秒単位で指定可能）経過したら音声を切り出して残りの処理に渡し，結果を表示する．以上の処理を繰り返す．

8. ま と め

本稿では，歌声と朗読音声の識別手法の高精度化を目的として，新たな音響特徴量の追加と，特徴量の統合方法の改良をした自動識別システムを開発し，評価を行った．まず，人間は歌声と朗読音声を識別する際に音素継続時間を利用し，長時間・短時間の音声に対し異なる手がかりを利用している可能性があることを確認した．そこで，従来識別に利用されてきた特徴量に，音素継続時間に関連した特徴量としてスペクトル変化量（アクセント）のピーク間隔を加え，その上で，各特徴量に対する識別時の重み付けが，入力音声長に応じて適切に変化するようにした．その適切な重みは，入力音声長毎に学習データから事前に求めた．実験の結果，1 秒の音声に対する識別精度は，従来法より 3.5%高い 90.2%となり，提案手法で適切に重み付けした有効性が確認された．その一方で，新たな特徴量であるスペクトル変化量のピーク間隔に関しては，単独での識別精度が 1 秒の音声に対して 70.4%と低い上

Vol.2012-MUS-94 No.13
Vol.2012-SLP-90 No.13
2012/2/4

に統合時の重みも小さく，今回の実験では有効性が確認されなかった．今後は，より現実的な場面を想定して，雑音を含む入力に対する自動識別にも取り組む必要がある．

謝辞 実験データの作成に必要なデータを提供して下さった，NTT コミュニケーション科学基礎研究所の大石氏に感謝する．本研究の一部は，科研費，GCOE の支援を受けた．

参 考 文 献

1) 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，電子情報通信学会技術研究報告. SP, 音声， Vol.97, No.560, pp.25–32 (1998).
2) 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，情報処理学会論文誌， Vol.47, No.6, pp.1822–1830 (2006).
3) 阿曽慎平，齋藤 毅，後藤真孝，糸山克寿，高橋 徹，尾形哲也，奥乃 博：F0・音韻長・パワー制御による歌声らしさ・話声らしさの変化の 評価，情報処理学会第 73 回全国大会, 2R-6 (2011).
4) Klapuri, A., Eronen, A. and Astola, J.: Analysis of the meter of acoustic musical signals, *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol.14, No.1, pp.342 – 355 (2006).
5) Scherer, K.R., Feldstein, S., Bond, R.N. and Rosenthal, R.: Vocal cues to deception: A comparative channel approach, *Journal of Psycholinguistic Research*, Vol.14, pp.409–425 (1985).
6) Friend, M. and Farrar, M.J.: A comparison of content-masking procedures for obtaining judgments of discrete affective states., *J Acoust Soc Am*, Vol.96, No.3, pp. 1283–90 (1994).
7) Erickson, D., Suzuki, T., Tanosaki, K., Yahiro, K., Haneishi, E. and Kishimoto, H.: Ah, how sweet the sound: Some acoustic characteristics of emotionally sung /ah/, *Intersinging* (2010).
8) 河原英紀：聴覚の情景分析が生み出した高品質 VOCODER:STRAIGHT，日本音響学会誌，Vol.54, No.7, pp.521–526 (1998).
9) Ryynanen, M. and Klapuri, A.: Modelling of Note Events for Singing Transcription, *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio* (2004).
10) Yegnanarayana, B., Sri Rama Murty, K. and Rajendran, S.: Analysis of stop consonants in Indian languages using excitation source information in speech signal, *ISCA ITRW Speech Analysis and Processing for Knowledge Discovery* (2008).
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，情報処理学会研究報告， Vol.2005, No.82, pp.7–12 (2005).