社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－MUS－61 (16)
2005／8／5

伴奏音抑制と高信頼度フレーム選択に基づく楽曲中の歌声の歌手名同定手法

藤原 弘将† 北原 鉄朗† 後藤 真孝‡
駒谷 和範† 尾形 哲也† 奥乃 博†

† 京都大学大学院 情報学研究科 知能情報学専攻 ‡ 産業技術総合研究所

本稿では，実世界の音楽音響信号に対する歌手名の同定手法について述べる．歌手名の同定を行う際に大きな問題となるのは，混在する伴奏音の影響である．本稿ではこの問題を解決するため，伴奏音抑制と高信頼度フレーム選択の手法を提案する．前者では，優勢なメロディの調波構造を抽出し再合成することで，伴奏音の影響を低減させることが出来る．後者は，歌声と非歌声を表わす2種類の混合正規分布を用いて，それぞれのフレームが信頼出来るか否かを判定するものである．実験の結果，本手法によって20歌手256曲に対して約93%の識別率が確認され，本手法を用いない場合と比較して誤り率が約65%削減された．

A Singer Identification Method for Singing Voice in Musical Pieces on the Basis of Accompaniment Sound Reduction and Reliable Frame Selection

HIROMASA FUJIHARA†, TETSURO KITAHARA†, MASATAKA GOTO‡,
KAZUNORI KOMATANI†, TETSUYA OGATA† and HIROSHI G. OKUNO†

† Dept. of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University
‡ National Institute of Advanced Industrial Science and Technology (AIST)

This paper describes a method for automatic singer identification from polyphonic musical audio signals. The main problem in automatically identifying singers is the negative influences caused by accompaniment sounds. To solve this problem, we developed two methods, *accompaniment sound reduction* and *reliable frame selection*. The former method makes it possible to reduce accompaniment sounds by extracting and resynthesizing hamonic components of the predominant melody. The latter method judges whether each frame of the obtained melody is reliable or not by using two Gaussian mixture models for vocal and non-vocal frames. Experimental results with 256 songs by 20 singers showed that our method was able to reduce 65% of classification errors, and achieved an accuracy of 93%.

1. はじめに

歌声は誰もが生まれながらに持つ最も基本的な楽器であり，多くのジャンルの音楽，特にポピュラー音楽において，重要な役割を果たしている．実際，多くの人は音楽を聞いたときに，その曲中の歌声を手掛かりにして曲名を判断するであろう．そのため，多くのCDショップ等では，音楽の分類にジャンルの情報に加えて歌手名(アーティスト名)を用いている．

このように歌声は音楽の重要な要素であるため，楽曲の歌手名に関する情報は音楽情報検索(MIR)に有用である．例えば，ある歌手の楽曲を探したい場合，歌手名(アーティスト名)のタグを利用することで，望みの楽曲を検索することが出来る．さらに，歌手の特性を表現する特徴量を記述し，それらの類似性を計算することが出来れば，声質に基づく楽曲検索の実現など，MIRにおいて重要な役割を果たすことが出来る．しかし，既存のMIRシステムのほとんどは，アーティスト名などのメタデータがすでに記述されていることを前提としている．そのため，そのようなメタデータが記述されていない楽曲は，アーティスト名をクエリーにしての検索は出来なかった．

本稿では，このような歌手名に基づく楽曲検索を実現するため，音楽音響信号から楽曲中の歌手名を同定する問題について検討する．歌手名の同定における最大の問題は，伴奏音の混在である．多くの楽曲には，歌声のみならず種々の伴奏音も含まれている．そのため，そのような音楽音響信号から抽出された歌声の特徴量は，伴奏音によって影響を受けている．実際，ケプストラム係数や線形予測係数 (LPC) など，音声認識でよく用いられる特徴量を音楽音響信号に対して計算すると，得られた特徴量は，歌声のみを表現するのではなく，歌声と伴奏音が混ざった状態を表現してしまう．音声情報処理の分野では，話し声に対する話者認識の研究が多く行われてきた[1),2)]が，ノイズのない話し声のみからなる信号を対象としたものが多いので，それらの研究での手法をそのまま歌手名の同定に用いることは難しい．

先行研究においては，Tsai ら[3)]はこの伴奏音混在の問題を指摘し，雑音下話者認識の手法[4)]を用いて解決しようとした．歌声と伴奏音は確率的に独立であるという仮定のもと，間奏部から推定された伴奏の確率モデルと，歌声が存在する区間から推定された伴奏と歌声の混ざった音のモデルを用いて，歌声のみのモデルを推定しようというものである．しかし，この仮定は常に満たされるわけではない．また，間奏部では歌声以外のリード楽器がメロディを演奏している場合が多く，推定された歌声のみのモデルの正当性に問題があった．その他の従来研究[5)~8)]では，伴奏音混在の問題は明示的に扱われてはいなかった．

この問題を解決するため，本稿では，**伴奏音抑制**と**高信頼度フレーム選択**の二つの手法を提案する．伴奏音抑制では，音響信号中のメロディの調波構造を抽出し正弦波重畳モデルを用いて再合成することで，メロディのみからなる音響信号を得る．高信頼度フレーム選択の手法により，歌声の特徴をよく保存していて，識別に相応しいフレームを選択し識別に用いる．ここでは，歌声を表わす確率モデルと，非歌声を表わす確率モデルの尤度比によって歌声の信頼度を判定する．

以下，第 2 章では，まず本研究での問題設定に取り組む上での課題と，その解決のためのアプローチについて述べる．第 3 章では，提案手法の実装の詳細について述べる．第 4 章では本手法の有効性を確かめるために評価実験を行い，第 5 章ではまとめを行う．

2. 伴奏音の影響に頑健な歌手名の同定手法

本稿では，与えられた音楽音響信号中の歌手名を同定する問題を扱う．歌手名の同定を実現する際に課題となるのは，歌声と同時に演奏される伴奏音の影響である．そのため，高精度の歌手名の同定を実現するには，このような伴奏音混在の問題を解決せねばならない．

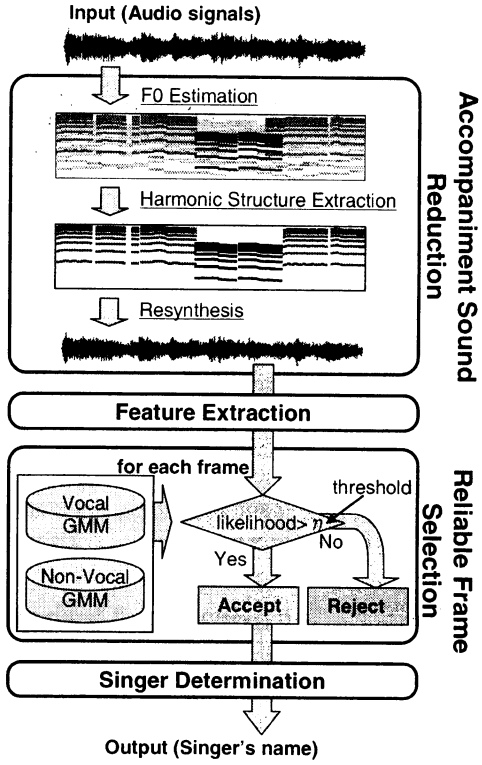


図 1　手法の概要

この問題の一つの解決策としてマルチコンディション学習，すなわち学習の際に伴奏音の影響を受けたデータを使用する方法が考えられる．実際，多くの従来研究ではこの方法が用いられていた[5)~8)]．しかし，伴奏音の性質は曲によって大きく異なるため，この方法は適切でないと考えられる．例えば，伴奏楽器にドラムス，ベース，ギターなどを含むフルバンドで演奏された楽曲と，ピアノの弾き語りの楽曲では，例え歌手が同じであっても，全体の音響的特性は大きく異なると考えられる．

本研究では，この伴奏音混在の問題を 2 つのアプローチで解決する．1 つは，伴奏音を入力信号から除去し，メロディのみの信号を生成することである．音響信号から伴奏音を除去することは困難ではあるが，入力信号をそのまま用いるのに比べると，伴奏音の影響を軽減できると期待できる．本研究では，これをメロディの調波構造抽出と正弦波重畳モデルを用いた再合成により実現し，**伴奏音抑制**と呼ぶ．ここで，メロディとは各時刻において最も優勢な音として定義する[9)]．さら

に，歌声の伴奏音による影響の受け方は時刻によって異なるため，影響の小さい部分のみを選択して同定に用いれば，よりロバストな歌手同定が期待できる．本研究では，これを**高信頼度フレーム選択**と呼び，歌声と非歌声を学習させた2つの混合ガウス分布（GMM）を用いて実現する．以下，それぞれの手法について詳説する．

2.1 伴奏音抑制

伴奏音抑制の手法は，以下の三つの処理からなる．

(1) 後藤の PreFEst[9] を用いて，メロディの基本周波数を推定する．

(2) 推定された基本周波数に基づき，メロディの調波構造を抽出する．

(3) 抽出された調波構造を，正弦波重畳モデルを用いて音響信号に再合成する．

以上の処理で，楽曲中のメロディのみの音響信号を得ることが出来る．

本手法により得られたメロディの音響信号は，間奏などの区間では(歌声でない)楽器音を含んでいる．そのため，歌声が存在する区間を前もって検出しておく必要があると考えられるが，歌声区間の検出は大変困難な問題である．本稿では，後述の高信頼度フレーム選択の手法を用いることでこの問題を回避する．

2.2 高信頼度フレーム選択

高信頼度フレーム選択では，歌声が存在する区間から抽出された特徴量で学習した歌声 GMM λ_{V} と，伴奏区間から抽出された特徴量で学習した非歌声 GMM λ_{N} の尤度比にを，各フレームの信頼度とする．

$\mathbf{x}$ を特徴ベクトルとしたとき，歌声 GMM の尤度 $p(\mathbf{x}|\lambda_{\mathrm{V}})$ はその特徴ベクトルの歌声らしさを表わし，非歌声 GMM の尤度 $p(\mathbf{x}|\lambda_{\mathrm{N}})$ はその特徴ベクトルの非歌声らしさを表わす．特徴ベクトルが伴奏音による影響をあまり受けていなかった場合，$p(\mathbf{x}|\lambda_{\mathrm{V}})$ は大きくなり，$p(\mathbf{x}|\lambda_{\mathrm{N}})$ は小さくなるため，信頼度は高くなる．すなわち，$\mathbf{x}$ の信頼度 $R(\mathbf{x})$ は，

$$R(\mathbf{x}) = \log p(\mathbf{x}|\lambda_{\mathrm{V}}) - \log p(\mathbf{x}|\lambda_{\mathrm{N}}) \qquad (1)$$

と表わされる．

本稿では，各曲の中から信頼度が上位の α% のフレームを識別に用いる．なぜなら，様々な楽曲に対して共通の閾値を決定すると，信頼度の高いフレームが少ない楽曲では，識別に十分な量の特徴ベクトルを確保出来ないからである．この手法により，歌声を含まないフレームの多くを排除出来る．つまり，高信頼度フレーム選択の手法を用いることで，歌声区間の検出を明示的に行う必要がなくなる．

3. 実　　装

この節では，本稿で述べる歌手名の同定システムの実装について述べる．前述のように，このシステムは，伴奏音抑制，特徴抽出，高信頼度フレーム選択，識別の4つの処理からなる．以下，それぞれの処理について詳説する．

3.1 前 処 理

入力音響信号を，モノラル化し，16 kHz にダウンサンプリングする．そして，フレーム幅 128 ms (2048 サンプル)，フレームシフト 10.0 ms (160 サンプル) で短時間フーリエ変換を行い，スペクトログラムを計算する．本研究では，窓関数にハミング窓を用いた．

3.2 伴奏音抑制

2.1 節で述べた手法を用いて，伴奏音を抑制する．

3.2.1 基本周波数推定

基本周波数推定 (F0 推定) には，後藤の PreFEst[9] を用いる．PreFEst は，制限された周波数帯域において最も優勢な調波構造を持つ F0 を推定する手法である．メロディは中高域の周波数帯域において最も優勢な調波構造を持つ場合が多いため，周波数帯域を適切に制限することで，メロディの F0 を推定することが出来る．

以下，PreFEst の概要を記す．以後，x は cent の単位で表わされる対数周波数軸上の周波数で，(t) は時間を表わすとする．パワースペクトル $\Psi_p^{(t)}(x)$ に対して，メロディの周波数成分の多くが通過するように設計された帯域通過フィルタを適用する．フィルタを通過後の周波数成分は $BPF(x)\Psi_p^{(t)}(x)$，と表わされる．ただし，$BPF(x)$ はフィルタの周波数応答である．以後の確率的処理を可能にするため，フィルタを通過後の周波数成分を確率密度関数 (PDF) として，以下のように表現する．

$$p_{\Psi}^{(t)}(x) = \frac{BPF(x)\Psi_p^{(t)}(x)}{\int_{-\infty}^{\infty} BPF(x)\Psi_p^{(t)}(x)dx} \qquad (2)$$

その後，周波数成分の PDF が，全ての可能な F0 に対応する音モデルの重みつき和からからなる確率モデル，

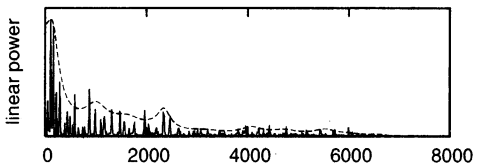
$$p(x|\theta^{(t)}) = \int_{\mathrm{Fl}}^{\mathrm{Fh}} w^{(t)}(F)p(x|F)dF, \qquad (3)$$

$$\theta^{(t)} = \{w^{(t)}(F)|\mathrm{Fl} \leq F \leq \mathrm{Fh}\} \qquad (4)$$

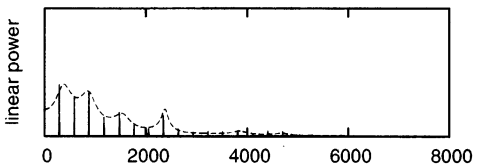
から生成されたと考える．ここで，$p(x|F)$ は，それぞれの F0 についての音モデルとし，Fh と Fl を取り得る F0 の上限と下限とする．また，$w^{(t)}(F)$ は音モデルの重みで，

$$\int_{\mathrm{Fh}_i}^{\mathrm{Fl}_i} w^{(t)}(F)dF = 1 \qquad (5)$$

を満たす．音モデルとは典型的な調波構造を表現した確率分布である．そして，EM アルゴリズムを用いて $w^{(t)}(F)$ を推定し，それを F0 の PDF と解釈する．最終



(a) An original spectrum and its envelope.



(b) An extracted spectrum and its envelope.

図 2　F0 推定，調波構造抽出の一例

的に，以下の式を用いて，最も優勢な基本周波数 $\overline{F^{(t)}}$ を決定する．

$$\overline{F^{(t)}} = \underset{F}{\operatorname{argmax}}\, w^{(t)}(F) \tag{6}$$

3.2.2 調波構造抽出

推定された F0 に基づき，メロディの調波構造の各倍音成分のパワーと位相を抽出する．それぞれの周波数成分の抽出の際には前後 $|r|$ セントずつの誤差を許容し，この範囲で最もパワーの大きなピークを抽出する．l 次倍音 ($l = 1, \ldots, 20$) の周波数 F_l，パワー A_l，位相 θ_l は，以下のように表わされる．

$$F_l = \underset{F}{\operatorname{argmax}} |S(F)| \quad (l\overline{F} \cdot (1 - 2^{\frac{r}{1200}}) \leq F \leq l\overline{F} \cdot (1 + 2^{\frac{r}{1200}})), \tag{7}$$

$$A_l = |S(F_l)|, \tag{8}$$

$$\theta_l = \arg S(F_l), \tag{9}$$

ここで，$S(F)$ はスペクトルを，$\overline{F}$ は PreFEst によって推定された F0 を表わす．本稿では，r の値として 20 を用いた．

図 2 に，F0 推定と調波構造抽出の例を示す．図 2 (a) は，伴奏音と歌声が混在したスペクトルとその包絡を表わし，図 2 (b) は，抽出された歌声の調波構造のスペクトルとその包絡を表わす．なお，スペクトルの包絡は線形予測分析を用いて計算されたものである．抽出後のスペクトル包絡は，元のスペクトル包絡と比較して，伴奏の影響が減っているいることがわかる．

3.2.3 再　合　成

抽出された調波構造を正弦波重畳モデルに基づき再合成することで，メロディの音響信号を得る．再合成された音響信号は，

$$s(t) = \sum_{l=1}^{L} A_l \cos(\omega_l t + \theta_l), \tag{10}$$

のように表わされる．ここで，A_l, θ_l, F_l はそれぞれ，l 次倍音のパワー，位相，周波数を表わし，t は時間を表わす．

3.3 特 徴 抽 出

再合成された音響信号から，特徴ベクトルを計算する．音声信号の個人性は，スペクトルの微細構造ではなく包絡に含まれていることが知られている[10]．音声認識，話者認識の分野では，スペクトル包絡を分離し特徴ベクトルを計算するための様々な手法が提案されている[11]．しかし，歌手名の同定の際に有効な特徴量についての考察は，今まで行われていなかった．そこで，本研究では，メル周波数ケプストラム係数 (MFCC)[12] と，線形予測メルケプストラム係数 (LPMCC)[13] の比較を行う．MFCC は音声信号に対してだけでなく，音楽音響信号に対しても一般的に使われている[14] 特徴量である．LPMCC は，線形予測分析 (LPC) によって得られる LPC スペクトルに対するメルケプストラム係数である．

3.3.1 メル周波数ケプストラム係数 (MFCC)

MFCC[12),14)] は，メル周波数軸上で計算されるケプストラム係数である．ケプストラム分析[11] とは，スペクトルの包絡と微細構造，つまり，声道特性と声帯振動を分離する手法である．ケプストラム係数は，対数パワースペクトルを離散コサイン変換することで計算される．スペクトル包絡はケプストラムの低次の係数に表現され，微細構造は高次の係数に表現される．メル周波数とは，人間の聴覚特性に適合した対数周波数軸である．MFCC の計算においては，まずメルフィルタバンク分析を行い，その後，対数をとり離散コサイン変換を行う．本稿では，12 次元 MFCC を用いた．

3.3.2 線形予測メルケプストラム係数 (LPMCC)

LPMCC は，線形予測スペクトル (LPC スペクトル) に対するメルケプストラム係数である．線形予測分析 (LPC)[15] とは，スペクトルの包絡と微細構造，つまり，声道特性と声帯振動を分離する手法の一つである．線形予測分析では，入力音響信号 $s(n)$ が与えられた場合に，ある時点での信号が，過去の一定期間の信号の 1 次結合で予測できると仮定する．予測値である $s_W(n)$ は，

$$s_W(n) = \sum_{i=1}^{p} \alpha_i s_W(n-i) + g(n), \tag{11}$$

のように与えられる．ここで，p は予測器の次数を表わし，予測係数である α_i は線形予測係数 (LPC) と呼ばれる．また，$g(n)$ はモデルの誤差を表わす．線形予

表 1　実験条件

	伴奏音抑制	フレーム選択
条件 i (ベースライン)	×	×
条件 ii	○	×
条件 iii	×	○
条件 iv (提案手法)	○	○

測係数は，誤差 $g(n)$ の二乗平均が最少となるように決定する．求められた線形予測係数を用いて，LPC スペクトルを，

$$|H(e^{j\omega})|^2 = \frac{1}{|1 - \sum_{i=2}^{p} a_i e^{-jiw}|^2} \qquad (12)$$

のように求めることが出来る．

LPC スペクトルに対するケプストラム分析は，基底の直交化の役割を果たしパターン認識の特徴量として有効であることが知られている．さらに LPMCC は，メル周波数軸を使用したことで，人間の聴覚特性に対する適合性という点で優れている．本研究では，LPC スペクトルから MFCC を計算することで，LPMCC を得た．

3.4 高信頼度フレーム選択

2.2 節で述べた手法に基づき，伴奏音の影響が少なく，信頼度が高いフレームを選択する．本稿では GMM の混合数として，64 混合を用いた．また，信頼度が高いと判定するフレームの割り合い α は，15%に設定した．

3.5 歌手名の決定

64 混合 GMM を用いて，歌手名を同定する．識別対象の歌手それぞれについて，GMM λ_s (s は歌手ラベル) を事前に学習しておく．$\mathbf{X} = \{\mathbf{x}_t | t = 1, \ldots, T\}$ を，高信頼度フレーム選択によって選ばれた特徴ベクトルとすると，歌手名は，以下の式に基づき決定される．

$$s = \underset{i}{\operatorname{argmax}} \frac{1}{T} \sum_{t=1}^{T} \log p(\mathbf{x}_t | \lambda_i) \qquad (13)$$

4. 実　　験

本節では，伴奏音抑制と高信頼度フレーム選択の手法の有効性を確認するために行った歌手名の同定実験について述べる．実験データとして，"RWC 研究用音楽データベース:ポピュラー音楽"[16] から選ばれたデータセットと，市販 CD から選ばれたデータセットの二種類を用いて実験を行った．

4.1 RWC 研究用音楽データベースを用いた実験

4.1.1 実 験 条 件

実験は表 1 の 4 つの条件で行われた．実験には，"RWC 研究用音楽データベース:ポピュラー"[16] から選ばれた 10 歌手 (男声 5 人，女声 5 人) による計 40 曲 (歌手 1 人当り 4 曲) を使用した．これらの楽曲の詳細を表 2 に記す．高信頼度フレーム選択の学習データには，表 3 に示される 16 歌手からなる 25 曲を用いた．

表 2　使用楽曲の内訳 (RWC 研究用音楽データベース)

	Name	Gender	Piece Number
a	西一男	M	012, 029, 036, 043
b	風戸ヒサヨシ	M	004, 011, 019, 024
c	森元康介	M	038, 039, 042, 044
d	井口慎也	M	082, 084, 088, 090
e	Jeff Manning	M	085, 087, 095, 098
f	吉井弘美	F	002, 017, 069, 075
g	緒方智美	F	007, 028, 052, 080
h	凛	F	014, 021, 050, 053
i	服部まきこ	F	065, 067, 068, 077
j	Betty	F	086, 092, 094, 096

表 3　高信頼度フレーム選択の学習に用いた楽曲.

Name	Gender	Piece Number
勝田真悟	M	027
波多江良徳	M	037
食原正機	M	032, 078
関谷洋	M	048, 049, 051
小澤克乃	M	015, 041
橋本まさし	M	056, 057
熊坂敏	M	047
オリケン	M	006
KONBU	F	013
市川えり	F	020
新田智子	F	026
鏑木朗子	F	055
飯島柚子	F	060
佐藤れいこ	F	063
松阪珠子	F	070
Donna Burke	F	081, 089, 091, 093, 097

これらの 16 歌手には，表 2 の 10 歌手は含まれていない．これらの 25 曲の各フレームに対して，歌声区間・非歌声区間の別をラベル付けし，歌声 GMM と非歌声 GMM を学習した．評価方法は，4-fold cross-validation 法を用いた．すなわち，各歌手の 4 曲の内 3 曲で学習し残りの 1 曲で評価する，という操作を，全ての楽曲が評価されるように 4 回繰り返した．1 回の評価では，まず，各歌手ごとに学習データを用いて GMM を学習し，識別では評価用楽曲各 1 曲に付き 1 つの歌手ラベルを出力した．特徴量は，3.3 節で述べた MFCC と LPMCC を両方用いて比較した．

4.1.2 結果と考察

図 3 に，RWC 研究用音楽データベースに対する実験結果を示す．伴奏音抑制と高信頼度フレーム選択のより，識別率が向上したことがわかる．また，伴奏音抑制と高信頼度フレーム選択を同時に用いた場合に，識別率が 55%から 95%と，大幅に向上していることがわかる．

図 4 に，特徴量として LPMCC を用いた場合の混同行列を示す．提案手法を用いることで，誤識別が減少

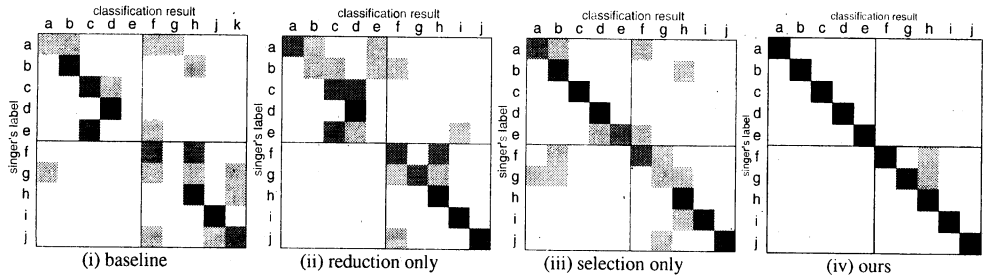


図 4　混同行列 (RWC 研究用音楽データベース). 特徴量は LPMCC を用いた. "reduction" と "selection" は, それぞれ伴奏音抑制と高信頼度フレーム選択を表わす. また, 各図中の点線は, 男女の境界を示す. 提案手法を用いることで, 誤識別が減少していく様子が見て取れる.

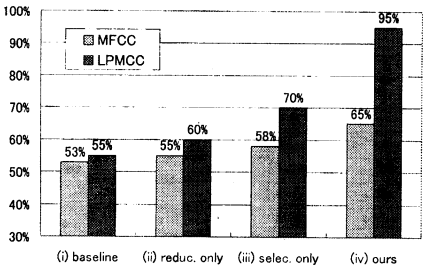


図 3　実験結果 (RWC 研究用音楽データベース). "reduc." と "selec." は, それぞれ伴奏音抑制と高信頼度フレーム選択を表わす.

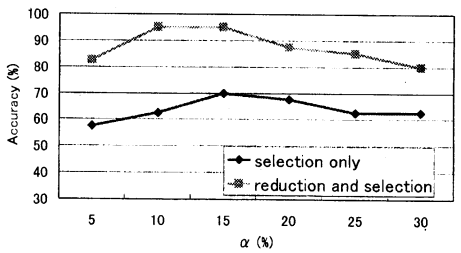


図 5　識別率の α の値による比較 (RWC 研究用音楽データベース)

し混同行列が対角行列に近づいていく様子が見て取れる. また, 図より, 伴奏音抑制を用いることで男女間の混同が減少していることがわかる. 伴奏音抑制を用いない場合 (i と iii) は, 伴奏音の影響によっていくつかの楽曲で男女間を混同しているが, 伴奏音抑制を用いた場合 (ii と iv) は伴奏音の影響が低減されたため, 性別を正確に判別出来るようになったと考えられる.

次に, MFCC を用いた場合と LPMCC を用いた場合とで結果を比較する. 全ての実験条件について, LPCMC のほうが, MFCC より識別率が高いことが確認できる. 特にこの傾向は, 伴奏音抑制と高信頼度フレーム選択を用いた場合に顕著である. LPMCC は, 歌手性をよく表わす特徴量であることが確認された.

信頼度が高いと判断されるフレームの割り合い α の値を様々に変化させた場合の識別率の変化を表 5 に示す. この図から, α を変化させたことによる識別率の変動は小さいと判断できる. また, 伴奏音抑制手法を用いた場合, α を 15%より大きくすることで識別率が大きく変化している. これは, 伴奏音抑制手法が伴奏音の影響を低減させたことで, 信頼性の高いフレームと低いフレームの差を際立たせたためと考えられる. そのため, α の値を上げすぎると信頼度の低いフレームをより多く選んでしまい識別率が低下した. 一方, 伴奏音抑制を用いない場合は, 各フレームの信頼度に大きな差がないため, α を大きくし多くのフレームを識別に用いても, 結果が大きく変化しなかったのだと思われる. しかし, その場合は, 伴奏音の影響により十分な識別率が得られていない.

4.2 市販 CD を用いた実験

4.2.1 実験条件

実験は, 前節での実験と同様に, 表 1 の 4 つの条件で行われた. 実験には, 市販 CD から選ばれた 20 歌手 (男声 8 人, 女声 12 人) からなる 246 曲を用いた. これらの楽曲の選定では, オリコンの 2004 年度年間アルバムチャートの上位から, 歌手が 1 人のアーティスト 20 アーティストほど選択し, それぞれに対して順位が最上位のアルバムに含まれる楽曲を使用した. 使用したアーティストを表 4 に示す. 高信頼度フレーム選択の学習データは, 4.1 節の実験と同様で, 表 3 に記される 16 歌手からなる 25 曲を使用した. 評価方法は, 3-fold cross-validation 法を用いた. すなわち, 各歌手の楽曲の 3 分の 2 を用いて学習し, 残りの楽曲を評価に用いるという操作を全ての楽曲が評価されるように 3 回繰り返した. その他の実験方法は, 4.1 節の実験と同様である.

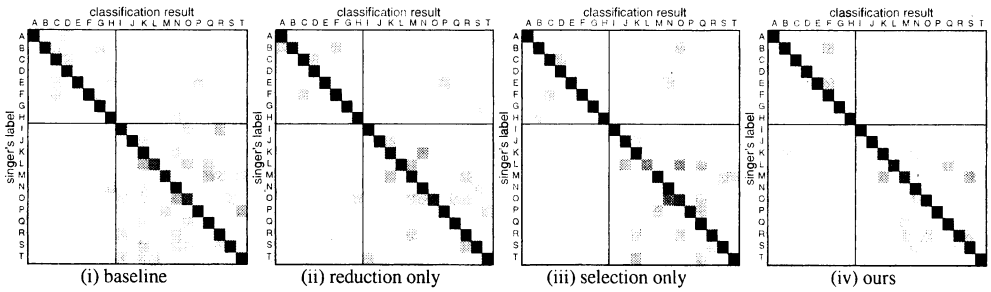


図 7　混同行列 (市販 CD)．特徴量は LPMCC を用いた．"reduction" と "selection" は，それぞれ伴奏音抑制と高信頼度フレーム選択を表わす．また，各図中の点線は，男女の境界を示す．提案手法により，誤識別が減少していくのが見て取れる．

表 4　使用楽曲の内訳 (市販 CD)

	Artist Name	Gender	Tracks
A	Asian Kung-fu generation	M	11
B	Bump of Chicken	M	10
C	平井堅	M	10
D	槇原敬之	M	12
E	森山直太朗	M	11
F	Mr.Children	M	12
G	ポルノグラフィティ	M	13
H	QUEEN	M	16
I	Aiko	F	13
J	Avril Lavigne	F	14
K	BoA	F	12
L	浜崎あゆみ	F	8
M	平原綾香	F	10
N	倉木麻衣	F	16
O	中島美嘉	F	13
P	大塚愛	F	11
Q	島谷ひとみ	F	15
R	柴咲コウ	F	12
S	宇多田ヒカル	F	15
T	矢井田瞳	F	12

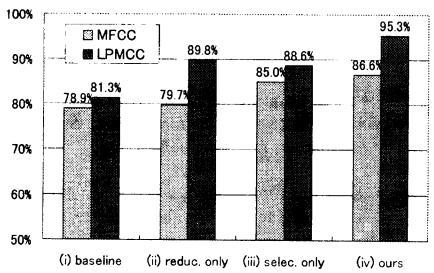


図 6　実験結果 (市販 CD)．"reduc." と "selec." は，それぞれ伴奏音抑制と高信頼度フレーム選択を表わす．

4.2.2 実験結果

表 6 に，実験結果を示す．伴奏音抑制か高信頼度フレーム選択のいずれか一方の手法を用いることで，約 8%の識別率の向上を確認した．さらに，両方の手法を同時に用いることで，識別率が約 12%向上すること確認した．本手法は，市販 CD に対しても有効に機能することがわかった．

図 7 に，特徴量として LPMCC を用いた場合の混同行列を示す．男声と比較して，女声の誤識別が多いことが読み取れる．これは，今回用いた 12 次元 LPMCC では，女声のスペクトル包絡と周波数微細構造を分離できなかったからであると考えられる．通常，話し声を対象にした場合，スペクトル推定におけるモデルのパラメータの次数は，女性や子どもの音声で，周波数微細構造の成分がスペクトル包絡の観測値に現われない程度に大きく設定する[13]．しかし，女性の歌声の F0 は時には 1000 Hz にも達するため，男声や F0 の低い女性の歌声を正確に表現できる程度にパラメータ次数を大きく取ると，F0 の高い女性の歌声では微細構造の成分がスペクトル包絡の観測値に表われてしまった．

市販 CD を用いた実験は，RWC 音楽データベースを用いた場合と比較して，ベースライン手法 (条件 i) の認識率が 15%程度高い．これは，市販 CD では各アルバムにおけるアーティストごとの特色を出すため，同じ楽器を使用するなど，同一アルバム内では全体の音質が均質になるように作られている場合が多いためと考えられる．そのため，伴奏音の影響を受けたままの特徴量で識別を行っても，81.3%とある程度高い識別率が得られた．Berenzweig ら[6] は，この現象を "アルバム・エフェクト" と呼び，歌手名の同定実験の識別率は，用いるデータセットの選び方に依存することを指摘している．一方，RWC 研究用音楽データベースは様々なジャンルにおける豊かなバリエーションを持つ楽曲を可能な限り大量に収録することを目標に構築されている[16] ため，同一歌手の楽曲でも使用楽器，曲調などが多様である．そのため，ベースライン手法では識別率が 55%程度しか達成出来なかった．しかし，実際には，複数のバンドに所属する歌手や，時期によって全体の

音質が大きく異なるアーティストも存在する．提案手法では，このような多様な楽曲を含むデータセットに対して高精度に歌手名を同定できた．

5. ま と め

本稿では，音楽音響信号から歌手名を同定する際の難しさとして，伴奏音混在の問題を指摘し，これを伴奏音抑制と高信頼度フレーム選択という2つの手法で解決した．前者は，メロディの調波構造を抽出・再合成することで伴奏音の影響の少ない信号を生成する手法で，後者は，あらかじめ学習した歌声モデルと非歌声モデルから，歌声らしさの高い特徴ベクトルだけを選ぶ手法である．これらの手法により，RWC研究用音楽データベース収録の10歌手40曲に対して55%から95%へ，市販CDに収録の20歌手246曲に対して81%から93%へ認識率を改善することができた．

本研究の意義を以下にまとめる．

- 歌手名を同定する際の難しさとして伴奏音混在の問題を明確化した．この問題は，一部の研究では指摘されていたものの，これまで明示的には扱われていなかった．
- 伴奏音混在の問題を解決する一手法として，メロディの調波構造の抽出と再合成に基づく伴奏音抑制を検討した．認識対象音の調波構造を抽出・再合成することで雑音の影響を除去する試みは音声認識の分野では行われているが[17]，この手法が歌手同定にも有効であることを確めたのは，本研究が初めてである．
- 伴奏音混在の問題のさらなる解決策として，歌声らしさの高い（伴奏楽器の影響が少ない）フレームのみを用いる高信頼度フレーム選択手法を提案した．これに類似した処理は先行研究でも行われていたが，歌声区間の検出に焦点が置かれ，信頼できるフレームのみを同定に用いるという観点には至っていなかった．本研究では，信頼度の観点からフレームを絞り込むことで，ロバストな歌手同定を実現した．

今後は，本手法を拡張することで，楽曲のボーカルの声質に基づく類似度を計算し，歌手の声質に基づく音楽情報検索を実現するための研究を進めていく．

謝辞 本研究の一部は，科学研究費補助金（基盤研究(A)，特定領域「情報学」），21世紀COEプログラム「知識社会基盤構築のための情報学拠点形成」の支援を受けた．また，本研究の実験において，「RWC研究用音楽データベース：ポピュラー音楽」(RWC-MDB-P-2001)[16]を使用した．最後に，ご討論いただいた吉井和佳氏，吉岡拓也氏(京都大学)に感謝する．

参 考 文 献

1) 松井知子: HMMによる話者認識，電子情報通信学会技術研究報告，SP95-111, No. 467, pp. 17–24 (1996).
2) 西田昌史: 音韻性を抑えた話者空間への射影による話者認識，電子情報通信学会論文誌，Vol. J85-D-II, No. 4, pp. 554–562 (2002).
3) Tsai, W.-H. and Wang, H.-M.: Automatic Detection and Tracking of Target Singer in Multi-Singer Music Recordings, *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, pp. 221–224 (2004).
4) Rose, R. C., Hofstetter, E. M. and Reynolds, D. A.: Integrated Models of Signal and Background with Application to Speaker Identification in Noise, *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 2, pp. 245–257 (1994).
5) Whitman, B., Flake, G. and Lawrence, S.: Artist Detection in Music with Minnowmatch, *Proceedings of the 2001 IEEE Workshop on Neural Networks for Signal Processing*, pp. 559–568 (2001).
6) Berenzweig, A. L., Ellis, D. P. W. and Lawrence, S.: Using Voice Segments to Improve Artist Classification of Music, *AES 22nd International Conference on Virtual, Synthetic, and Entertainment Audio* (2002).
7) Kim, Y.E. and Whitman, B.: Singer identificatin in popular music recordings using voice coding features, *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR2002)*, pp. 164–169 (2002).
8) Zhang, T.: Automatic Singer Identification, *Proceedings of IEEE International Conference on Multimedia & Expo (ICME 2003)*, Vol. I, pp. 33–36 (2003).
9) Goto, M.: 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).
10) 古井貞熙: 音声波に含まれる個人性情報の研究，博士論文，東京大学 (1978).
11) Picone, J.: Signal Modeling Techniques In Speech Recognition, *IEEE Proceedings*, Vol. 81, No. 9, pp. 1215–1247 (1993).
12) Davis, S. B. and Mermelstein, P.: Comparison of parametric representation for monosyllabic word recognition, *IEEE Transactions on Acoustic, Speech and Signal Processing*, Vol. 28, No. 4, pp. 357–366 (1980).
13) 今井聖: 音声信号処理 音声の性質と聴覚の特性を考慮した信号処理，森北出版株式会社 (1996).
14) Logan, B.: Mel frequency cepstral coefficients for music modelling, *Proceedings of the International Symposium on Music Information Retrieval (ISMIR 2000)*, pp. 23–25 (2000).
15) Shikano, K.: Evaluation of LPC spectral matching measures for phonetic unit recognition, Technical Report CMU-CS-96-108, CMU, Computer Science Department (1986).
16) 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).
17) Nakatani, T. and Okuno, H. G.: Harmonic Sound Stream Segregation Using Localization and Its Application to Speech Stream Segregation, *Speech Communcations*, Vol. 27, pp. 209–222 (1999).