歌声の調波構造抽出を用いた歌手名の同定

3R-8

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 工学部情報学科 ‡ 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 (AIST)

1. はじめに

本稿では，市販 CD 等の実世界の音楽音響信号に対して歌手名を自動的に同定する手法について述べる．歌声は多くのジャンルの楽曲で中心的な役割を果たしている．そのため，歌声の理解は計算機による音楽理解のなかでも最も重要な課題の 1 つであり，声質による音楽検索や，音楽データベースの自動インデキシングなど幅広い用途で有用となる．

実世界の音楽音響信号に対する歌手名の同定で問題となるのは，歌声と同時に演奏される種々の伴奏楽器の影響である．これらの影響により，歌手が異なるが伴奏楽器の音色が似た楽曲を同一の歌手と判断する等の誤認識が発生する．実際，Berenzweig[1] らの手法では，伴奏を含む音響信号に対してメル周波数ケプストラム係数 (MFCC) を計算していたものの，学習と認識に収録アルバムの異なる楽曲が用いられると，認識率が低下していた．また Tsai[2] らの手法では，伴奏音が確率的に定常と仮定し，歌声の確率モデルを，伴奏を含む歌声のモデルと間奏部のモデルから推定することを試みたものの，実際の楽曲では伴奏音が定常ではないなどの問題があるため限界があった．このように，従来研究では伴奏音混在の問題は十分には解決されていない．

本研究では，伴奏を含む音響信号から歌声のみを分離することで，この問題に対処する．まず伴奏音を含む信号において歌声の基本周波数 (F0) を推定し，その F0 に基づいて調波構造を抽出する．そして，抽出された調波構造を正弦波重畳モデルを用いて再合成することで伴奏音の影響が軽減された音響信号を得る．その音響信号に対して，線形予測分析 (LPC) ケプストラム係数を計算し，学習・識別を行う．

2. 歌手名の同定手法

本研究では，歌手名が付与された歌声の特徴ベクトルのデータベースに基づいて，提示された楽曲中の歌手名を同定する．ただし，本稿では歌声が存在する区間はあらかじめ付与してあるとし，一つの楽曲に複数の歌手が存在する楽曲は扱わない．

歌手名の同定処理は，1. 歌手の音響信号の分離，2. 特徴抽出，3. 学習・識別，の手順で行われる．以下に，それぞれの処理について詳細を述べる．なお，用いる音響信号は 16kHz でサンプリングしてある．また，以下の処理は，シフト幅 10ms，幅 128ms のハミング窓を用いた短時間フーリエ変換によりスペクトルを計算した後に行う．

2.1 歌声の音響信号の分離

伴奏音の影響を軽減するために，歌声の調波構造を抽出し，再合成することで歌声が分離された音響信号を得る．これらの処理は，以下のように行う．

2.1.1 F0 推定

後藤の PreFEst [3] を用い，歌声の F0 推定をする．PreFEst は意図的に制限された周波数帯域において最も優勢な調波構造の F0 を EM 法を用いて推定する手法であり，帯域の制限を適切に設定することで歌声の F0 を推定することが出来る．この手法では，間奏部などで歌声以外の優勢な楽器音の F0 を推定するが，本稿では歌声が存在する区間を手動で切り出しているので，この問題は生じないものとする．

2.1.2 調波構造の抽出

上の処理で求められた F0 に基づいて，基本周波数成分と高調波成分のパワーを 20 次倍音まで抽出する．各成分の抽出においては，前後 20cent 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．

2.1.3 再合成

抽出された調波構造を正弦波重畳モデルを用いて再合成する．調波構造の l 次倍音のパワー，角周波数および位相を，A_l，ω_l，θ_l とすると，再合成された波形は，

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

で表わされる．ただし，t は時刻を表わす．

2.2 特徴抽出

上記の処理で得られた歌声の音響信号に対して LPC 分析を行い，LPC 分析によって得られたスペクトル包絡 (LPC スペクトル) に対してケプストラム分析を行うことで特徴量を計算する．

音声信号に含まれる個人性の情報は，スペクトルの微細構造ではなく包絡に多く含まれていると言われている [4]．LPC 分析は，音声信号からスペクトル包絡を分離する手法の一つであり，広く用いられている．しかし，LPC 分析により得られる LPC 係数は基底が直交化されておらず，パターン認識には不向きである．これを直交化する手段として，LPC スペクトルに対してケプストラム分析を行うことが有効である．本稿では，ケプストラム分析の手法として，MFCC を用いる．

以下に LPC 分析，及び MFCC について説明する．

2.2.1 LPC 分析

LPC 分析では，音声信号が式 (2) のような全極形の伝達関数を持つ調音フィルタの出力であることを仮定する．

$$H(z) = \frac{b_0}{a_0 - a_1 z^{-1} - a_2 z^{-2} - \cdots - a_p z^{-p}} \qquad (2)$$

そして，伝達関数を規定するパラメータである線形予測係数 $a_i (i = 1, \cdots, p)$ を決定することで，音声信号からスペクトル包絡を抽出する．抽出された LPC スペクトルは式 (2) を用いて $|H(e^{jw})|$ で表わされる．ただし，p は分析次数で，本稿では 25 とする．図 1 に調波構造抽出前後のスペクトルと LPC スペクトルを示す．

2.2.2 MFCC

MFCC は音の高低に関する人間の感覚尺度であるメル周波数軸上でのケプストラム分析手法である．まずメル

Singer Identification Using the Harmonic Structure of Singer's Voice: *Hiromasa Fujihara (Kyoto Univ.), Tetsuro Kitahara (Kyoto Univ.), Masataka Goto (AIST), Kazunori Komatani (Kyoto Univ.), Tetsuya Ogata (Kyoto Univ.), and Hiroshi G. Okuno (Kyoto Univ.)*

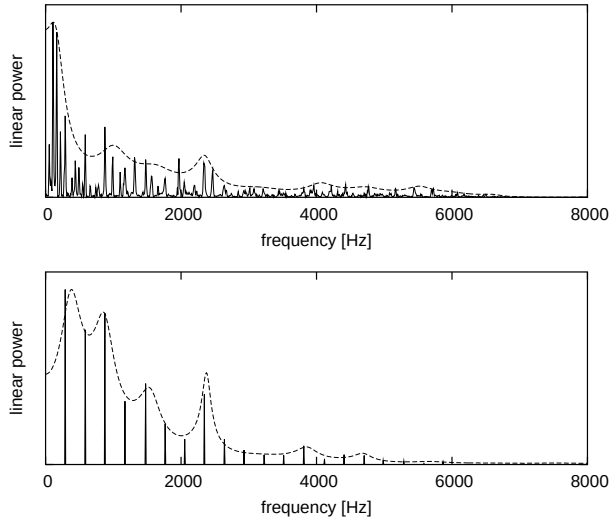

図 1: 調波構造抽出・スペクトル包絡分離の様子 : 上図が元のスペクトルと LPC スペクトル，下図が抽出後の調波構造と LPC スペクトル

表 1: 使用楽曲の内訳

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

周波数軸上で L 個の等間隔の三角窓を配置し，フィルタバンク分析を行う．次に，フィルタバンク分析により得られたそれぞれの帯域におけるパワーの対数を離散コサイン変換して MFCC を求める．本稿では $L = 15$，MFCC の次数は 12 とする．

2.3 学習・識別

学習には 64 混合 GMM を用いる．識別は楽曲ごとに行い．フレーム毎の対数尤度の総和が最大となる歌手のラベルを結果として出力する．

3. 評価実験

3.1 実験条件

本手法の有効性を確認するため，楽曲に対して歌声成分を分離せずそのまま MFCC を用いる特徴抽出手法 (比較手法) と前述の調波構造抽出に基づく特徴抽出手法 (本手法) の比較実験を行った．また，参考として，F0 推定の精度が十分に高い場合を評価するため，F0 推定にのみ伴奏なしの音響信号を用いた場合も行った．実験データには「RWC 研究用音楽データベース:ポピュラー音楽」[5] 中の楽曲から 10 歌手 (男声 5 人，女声 5 人) に対して各 4 曲ずつ，合計 40 曲を用いた．使用した楽曲を表 1 に示す．ただし，実験データはボーカルの存在する区間 60 秒 (6000 サンプル) を手作業で切り出して使用している．評価は 4-fold cross-validation により行った．

表 2: 実験結果

	比較手法	本手法	参考 (注)
a	3/4	3/4	3/4
b	4/4	4/4	4/4
c	3/4	4/4	4/4
d	4/4	4/4	4/4
e	3/4	2/4	3/4
f	2/4	2/4	3/4
g	0/4	2/4	2/4
h	4/4	4/4	4/4
i	4/4	4/4	4/4
j	3/4	4/4	4/4
Total	75.0%	82.5%	87.5%

(注) F0 推定にのみ伴奏なしの音響信号を用いた場合

3.2 実験結果と考察

結果を表 2 に示す．本手法を用いることにより認識率が 7.5%上昇しており，一定の効果が確認出来た．

歌手ごとの認識結果を見ると，歌手 g は比較手法では 1 曲も正解出来なかったのに対して，本手法では 2 曲正解であった．歌手 g で用いた楽曲のジャンルが，ピアノの弾き語り，R&B，ニューミュージック，J-POP と大きく異なり，伴奏楽器の種類や音質に差異があった．このため，比較手法では伴奏音の音質が似た他の歌手の楽曲と混同し同定に失敗した一方，本手法を用いることで伴奏音の影響が軽減され，より正確な同定が行われている．

反対に歌手 e では本手法により認識誤りが増加した．F0 推定の誤りにより，歌声の調波構造を抽出に失敗したためと考えられる．これは，F0 推定に伴奏なしの音響信号を用いた場合には認識誤りは増加していないことからも確かめられる．実際，この歌手の楽曲は，F0 が他の優勢な伴奏音としばしば混同されていた．今後，F0 推定の精度を向上させるためには，調波構造の優勢さだけでなく，歌声かどうかの判別処理を組み込む必要がある．

4. おわりに

本稿では，楽曲中の歌手名の同定手法について述べた．伴奏音が歌手名の同定に悪影響を及ぼすことを実際に確認し，歌声の調波構造を抽出することで，伴奏音の影響を軽減することを可能にした．今後は，調波構造抽出処理の改善により同定精度を向上させることを目指すとともに，楽曲中の歌声が存在する区間を検出する手法について研究を進めていく予定である．本研究の一部は科研費，21 世紀 COE の支援を受けた．

参考文献

[1] A. Brerenzweig, D. P. W. Ellis, and S. Lawrence, “Using Voice Segments to Improve Artist Classification of Music”, *AES 22nd Int’l Conf. on Virtual, Synthetic, and Entertainment Audio*, 2002.

[2] Wei-Ho Tsai and Hsin-Min Wang, “Automatic Detection and Tracking of Target Singer in Multi-Singer Music Recordings”, *Proc. ICASSP*, Vol.IV., pp.221–224, 2004.

[3] Masataka Goto, “A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals”, *Speech Comm.*, 43, pp 311–329, 2004.

[4] 古井貞熙, “音声波に含まれる個人性情報の研究”, 東京大学博士論文, 1978.

[5] 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.