歌声のゆらぎに着目した歌声強調に基づく音楽音響信号のメロディライン推定*

◎橘秀幸 (東大院情報理工), 小野拓磨 (東大工), 小野順貴, 嵯峨山茂樹 (東大院情報理工)

1 はじめに

我々が日常的に耳にする楽曲の多くには歌声によるメロディが含まれている。歌声によるメロディは音楽の中で特にキャッチーなパートであり，楽曲からメロディラインの情報を抽出することができれば，音楽情報検索などにおいて有効であると考えられる。

ところで，人間の場合，メロディラインを知覚するのは困難ではないが，伴奏音が混合された信号からメロディに関する情報を抽出するのは計算機にとっては困難な問題である。音楽信号から計算機で自動的にメロディラインの情報を抽出する手法として，PreFEst[1] を始めとし，[2, 3] などが知られている。また，音楽情報検索の世界的なコンテストである MIREX[4] でも，メロディラインの推定に関するタスクがあり，いくつかの手法が提案されている。

本稿では，歌声には基本周波数や振幅にヴィブラート的な変動のある場合が多いことに着目し，これらのヴィブラート的な成分を強調する手法として我々がこれまでに提案した方法 [5, 6] を利用してまず歌声成分を強調し，次に動的計画法による経路探索によって歌声のメロディラインの推定を行う手法を提案する。

2 歌声のゆらぎを利用した歌声強調 [5, 6]

歌声にはヴィブラートなどに由来するスペクトルゆらぎがある。このため，時間周波数解析をしたときに，歌声のスペクトルの形状は時間周波数分解能によって大きく異なるという性質がある。具体的には，時間分解能を高く（周波数分解能を低く）した場合は，歌声はほとんど定常的な音と見なせるため，歌声のスペクトルは時間軸方向に滑らかな成分として現れるのに対し，その逆に時間分解能を低く（周波数分解能を高く）した場合は，歌声の非定常性のためスペクトルは時間軸方向には滑らかにならないことと，周波数分解能が高いため微細な変動でも検出されやすくなることから，周波数軸方向に滑らかな成分として現れる。

このような，時間軸方向や周波数軸方向に滑らかといった性質は，調波打楽器音分離 (Harmonic/Percussive Sound Separation: HPSS)[7] という手法によって検出が可能であり，時間軸方向成分，周波数軸方向成分のそれぞれの成分に分離することができる。

これより，入力とする音楽を 2 通りの時間周波数分解能によって時間周波数表現し，そのそれぞれで HPSS を行うことで，

a) 前者の時間周波数分解能で，時間方向成分（コード楽器音+歌声）と周波数方向成分（打楽器音）を分離
b) 後者の時間周波数分解能で，時間方向成分（コード楽器音）と周波数方向成分（歌声+打楽器音）を分離

の 2 通りの分離が可能であり，両者の処理を段階的（または並列的）に行うことで歌声成分が強調された信号を得ることができる。

3 動的計画法によるメロディラインの探索

3.1 経路探索モデル

前節によって得られた歌声強調信号から歌声のメロディラインを動的計画法により探索する。

いま，入力信号を時間周波数解析して得た短時間スペクトル $\mathrm{s}_t = \{s_{t,x}\}$ の時系列を $S_n = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$ とし，その背後にあるメロディの音高（対数周波数）の時系列（メロディライン）を $X_n = \{x_1, x_2, \ldots, x_n\}$ とする。本稿では，S_n が得られたときに，以下のように事後確率を最大にするような最も確からしいメロディライン $\hat{X}_n$ を求めることを考える。すなわち，

$$\hat{X}_n = \operatorname*{argmax}_{X_n} p(X_n|S_n) \quad (1)$$

$$\propto \operatorname*{argmax}_{X_n} p(S_n|X_n)p(X_n) \quad (2)$$

ここで，時刻 t の音高 x_t は直前の音高 x_{t-1} のみに依存し，スペクトル $\mathbf{s}_t$ は音高 x_t のみに依存すると仮定すると，式 (2) は

$$p(S_n|X_n)p(X_n) \propto \prod_{t=1}^{n} p(\mathbf{s}_t|x_t) \prod_{t=1}^{n} p(x_t|x_{t-1})p(x_0) \quad (3)$$

$$= \prod_{t=1}^{n} \{p(\mathbf{s}_t|x_t)p(x_t|x_{t-1})\}\, p(x_0) \quad (4)$$

と展開することができる。これにより，最適なメロディライン $\hat{X}_n$ を求める問題は，式 (4) の各時刻に関する確率分布 $p(\mathbf{s}_t|x_t)p(x_t|x_{t-1})$ を x_{t-1} から x_t への遷移スコアとみなした経路探索問題に帰着する。

この問題は，動的計画法と呼ばれるアルゴリズムによって効率的に解くことができる。具体的には，各時刻の各音高 x_t に関して，

$$\ln p(x_t, X_{t-1}|S_t) = \ln p(x_{t-1}, X_{t-2}|S_{t-1}) + \ln p(\mathbf{s}_t|x_t) + \ln p(x_t|x_{t-1}) \quad (5)$$

を最大にするような x_{t-1} を記憶しておき，時刻 n にて $p(x_n, X_{n-1}|S_n)$ が最大となっているような $\hat{x}_n$ を起点として，$\hat{x}_{n-1}, \hat{x}_{n-2}, \cdots, \hat{x}_1$ を後ろ向きに辿っていくことで，各時刻の音高の推定値を得る。ここで用いる $p(\mathbf{s}_t|x_t)$ と $p(x_t|x_{t-1})$ については，それぞれ以下のようなモデルを仮定した。

3.2 スペクトルからの音高推定

短時間スペクトル $s_t(\xi)$ と音色モデル $q(\xi)$ との相関関数を考えると，その値が大きいような音高に，真の音高 x_t が存在している可能性が高いと考えられる。そこで，本稿では両者の相関関数をそのまま各時刻での音高の尤度 $p(\mathbf{s}_t|x_t)$ とみなした。すなわち，

$$p(\mathbf{s}_t|x_t) \propto \sum_{\xi} s_t(x_t + \xi)q(\xi). \quad (6)$$

ただし ξ は対数周波数を離散化したもので，$q(\xi)$ は調波構造を持つような音色モデルであり，第 n 倍音の強度が基音の強度の $1/n$ であるような音色を仮定した。

---

*Melody Line Estimation in Music Audio Signals using Singing Voice Enhancement Method Based on Vocal Spectral Fluctuation, by TACHIBANA Hideyuki, ONO Takuma, ONO Nobutaka and SAGAYAMA Shigeki (The University of Tokyo).

3.3 メロディラインの遷移モデル

音高 x_t のとりうる値の集合として，一般的な歌で出現頻度が高いと考えられるテノールの低音からアルトの高音までを想定し，A2 (110Hz) 以上 E5 (660Hz) 未満の区間で 25 セント刻みの 124 状態を仮定した。本稿では，音高差が小さいほど遷移しやすいというモデルを考え，この遷移確率をガウス関数で与えた。すなわち，

$$p(x_t|x_{t-1}) \propto \exp -\frac{(x_t - x_{t-1})^2}{2\sigma^2}. \qquad (7)$$

4 実音楽信号による評価実験

4.1 実験条件

実験には，MIREX の Audio Melody Extraction にて参考用のデータとして公開されている LabROSA データセット [8] を使用した。このデータセットは 13 曲からなり，各楽曲 30[s] 程度の長さで，サンプリング周波数 44.1kHz のモノラル信号である。各オーディオデータにはメロディラインの正解データが付属している。楽曲 1 から楽曲 9 までは歌声がメロディの楽曲であり，実験にはこれら 9 曲を使用した。なお，本稿では計算量を削減するためこれらを 16kHz にリサンプリングして使用した。

2 節の歌声強調では直列多重 HPSS[6] を採用し，HPSS のフレーム長は長い方を 4096 点，短い方を 512 点とした。分離に用いる定数のセットは，長い方では $(w_h, w_p) = (1.01, 1.00)$，短い方では $(1.00, 1.01)$ とすることで，非歌声成分の抑圧を弱めにした。3 節のトラッキングに用いるスペクトルの時系列は，ガボール関数をマザーウェーブレットとしたウェーブレット変換により求め，周波数分解能は 25 セント，時間分解能は 10[ms] とした。経路探索の初期条件 $p(x_0)$ には，中音域 $x_0 =$ E4 付近に大きめの確率を与えた。式 (7) での σ は 50 セントとした。

4.2 実験方法

これらの条件下で，各楽曲について 2 節の歌声強調，3 節の経路探索の手法を適用してメロディラインを推定し，得られた音高の推定値 $\hat{X}_n$ と正解データとを比較して，推定値と正解の差が半音以内であれば正解とみなして正解率を求めた。

また，本手法は休符を考慮していない手法であるため，休符を含む全区間に関する正解率のほかに，休符区間以外での正解率も求めた。また，2 節の歌声強調の効果を検証するため，歌声強調を行わずに経路探索のみを行った場合についても同様に実験し，比較した。

4.3 実験結果・考察

楽曲 1 に関する，正解のメロディラインと推定されたメロディラインを，歌声強調あり，なしの場合それぞれについて，Fig. 1 に示す。Fig. 1(下) では，1 オクターブ下を追跡している箇所が見られるが，これは低音の伴奏が抑圧されていなかったために，メロディラインが音高推定の段階で低音域に加算されたためと考えられる。これに対し Fig. 1(上) では，そのような伴奏が抑圧されているため，歌声のラインが追跡されやすくなっている。

また，その他の楽曲に関しても，歌声強調を行わない場合は伴奏などに捕捉されやすくなるため正解率があまり高くなかったのに対し，歌声強調を行った場合は妨害する伴奏が抑圧されるため正解率の大幅な向上が認められた。各楽曲についての正解率を Table 1 に示す。

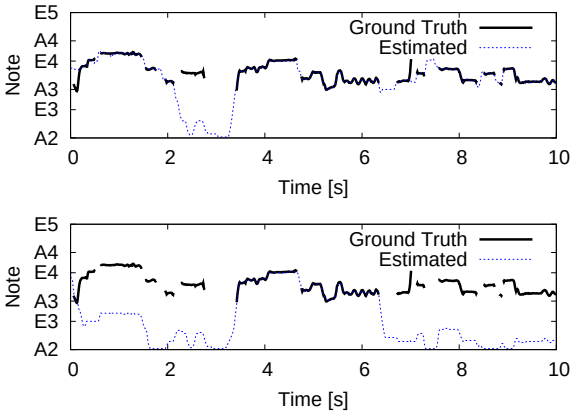

Fig. 1 楽曲 1 の冒頭 10 秒間での，推定したメロディラインと正解との比較。上：歌声強調あり，下：歌声強調なし

Table 1 各楽曲での正解率

	歌声強調なし		歌声強調あり	
楽曲	全正解率	休符を除外した正解率	全正解率	休符を除外した正解率
1	22%	34%	52%	79%
2	36%	67%	55%	81%
3	28%	46%	45%	76%
4	29%	44%	47%	71%
5	58%	88%	60%	93%
6	13%	26%	35%	72%
7	46%	68%	51%	76%
8	10%	15%	48%	71%
9	54%	72%	70%	93%

5 結論

本稿では，音楽音響信号に対し，歌声のスペクトルゆらぎなどに着目した歌声強調手法を適用し，次にスペクトルと整合的かつ連続的なラインを動的計画法によって探索することでメロディラインを検出する手法を提案した。

実験の結果，歌声強調を前処理として適用することにより，直接メロディを探索する場合と比較して，大幅な性能の改善が認められた。なお，データセットが異なるものの従来法ではおよそ 80%程度の正解率に達しており，本手法は更なる改良を要する。今後，歌声強調信号と同時に得られる歌声抑圧信号を積極的に利用することで，歌声強調信号に残存した伴奏を追跡することを防止することの検討や，状態の集合に休符モデルを組み込むことで全体の正解率を向上させることなどが課題となる。

参考文献

[1] Goto, *Speech Communication,* Vol.43, pp.311–329, 2004.

[2] Poliner *et al.*, *IEEE Trans. ASLP,* Vol.15, No.4, pp.1247–1256, 2007.

[3] Li, Wang, *IEEE Trans. ASLP,* Vol.15, No.4, pp.1475–1487, 2007.

[4] Downie, *Acoust. Sci. & Tech.,* Vol.29, No.4, pp.247–255, 2008.

[5] 橘他, 音講論 (春), pp.853–854, 2009.

[6] 橘他, 情処研報, MUS-81, 2009.

[7] Ono *et al.*, *Proc. ISMIR,* pp.139–144, 2008.

[8] `http://labrosa.ee.columbia.edu/projects/melody/`