Vol.2013-MUS-100 No.18
2013/9/1

鼻歌検索システムのための楽曲からのボーカルメロディ推定

角尾 衣未留[1,a)]　井上 晃[1,b)]　西口 正之[1,c)]

概要：鼻歌検索の参照メロディーデータとして用いることを主目的にしたボーカルメロディ推定手法について提案する．ボーカル音声は 3 つの特徴を持っており，ほんのわずかな時間でもピッチが変動していること，他の楽器音と異なる定位位置にミックスされていること，連続性をもっていること，が挙げられる．我々はこの特徴を用いて 3 ステップでのボーカルメロディ推定を行う．まず，周波数軸方向にローパスフィルタ処理を行うことでボーカルのスペクトルを強調する．次にミキシングの定位情報を用いてボーカルのスペクトルをさらに強調する．最後に連続性を考慮した新しい動的計画法を用いて最適なピッチ軌跡を推定する．推定されたボーカルメロディは動的計画法をベースとした鼻歌検索の参照データとして用いられ，実験によって従来のボーカルメロディ推定手法を用いた場合より高い性能が示された．

1. はじめに

大量の楽曲がインターネットで利用され個人で所有する昨今，その検索手段はアーティスト名や曲名などの単語によるものが主であり，それ以外に未だ成熟した方法は少ない．個人で所有する楽曲全てにメタ情報を付与することは容易ではなく，そのメタ情報を検索時に必ず思い出すとも限らない．メタ情報以外による検索手段は多くのアプローチが検討されているが，鼻歌検索はその中でも有力な手段の一つである．

鼻歌検索は比較的長い間研究され続けている．最も初期の研究は楽譜情報や MIDI 情報を検索対象としていた．鼻歌や歌唱による入力音声の不正確さのため，鼻歌検索システムに動的計画法を採用したり [1], [2]，メロディの上下の変化を U/D/S の 3 種類に量子化するなどして [3], [4]，ロバストなマッチングを行う研究が多い．しかし，レコーディングされた楽曲に比べて利用できる楽譜は多くなく，検索対象としては音響信号の方が需要が高い．

音響信号を対象とした場合はシンボリックな楽譜情報を扱うより難易度が高くなる．そのため，1 つの解決方法として中間的な情報を用いるやり方がある．西村らはメロディらしさのパターンを抽出し，抽出した全てのパターンを用いて入力クエリとマッチングする方法を提案している [5]．また，Song らは解析セグメントごとに複数音高の候補のセットを抽出する方法を提案している [6]．

一方，音響信号から直接メロディ情報を抽出して MIDI 情報と同様に扱うことによって鼻歌検索を行う方法も考えられる．メロディの推定自体が難しい問題であり，多くの研究者がこの研究に取り組んでいる．ほとんどのメロディ推定手法は複数のピッチ候補を抽出しそれらの中から最適なメロディラインを選択する．後藤は複数のトラッカを用いて有力な候補から主メロディを抽出する手法を提案し [7]，Li ら隠れマルコフモデル（HMM）と Viterbi サーチを用いて最適なメロディを推定している [8]．また，必ずしもメロディが他の楽器音と比べて目立つとは限らないため，橘らはボーカル音声をパワースペクトルの時間方向と周波数方向の変化量の違いを用いて抽出し [9]，Hsu らはピッチの大まかな変化をピッチトレンドとして用いることでメロディ推定精度を向上させ [10]，Salamon らはピッチトレンドと類似の方法で，ピッチ軌跡とピッチ平均の計算との組み合わせにより優れた推定性能を示した [11]．

本論文では鼻歌検索システムの参照データとして用いることを主目的としたメロディ情報推定手法について提案する．第 2 節でメロディ推定手法について説明する．まず，ボーカル音声の性質を用いてスペクトログラム中のボーカル音声部を強調し，ピッチ軌跡を抽出し，新しく定式化した動的計画法を用いてその中から最適なメロディラインを選択する．第 3 節で推定されたメロディ情報が鼻歌検索においてどのように用いられるかを簡単に説明し，第 4 節で実際の音響信号に対して適用することで提案手法の有用性について評価する．最後に，第 5 節でまとめと今後の展望について述べる．

[1] ソニー株式会社
Sony, Minato, Tokyo 108–0075, Japan
[a)] Emiru.Tsunoo@jp.sony.com
[b)] AkiraB.Inoue@jp.sony.com
[c)] Masayuki.Nishiguchi@jp.sony.com

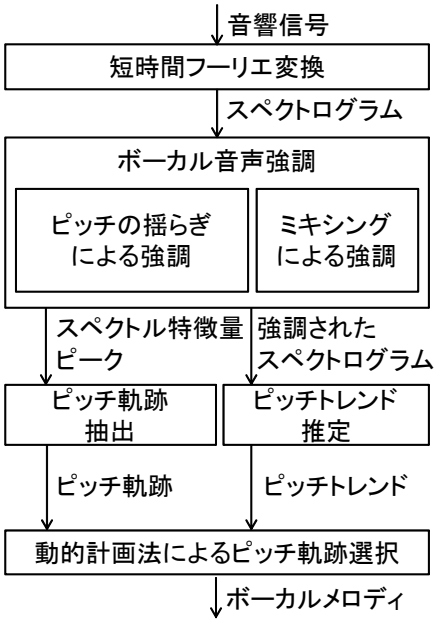

図 1　メロディ推定のフロー図

2. ボーカルメロディ推定

2.1 ボーカル音声の性質

一概にメロディには歌声のみでなく，ジャズのトランペットなど，様々な楽器によるものが考えられるが，鼻歌検索においてはボーカルの歌声によるメロディが検索対象となる場合が最も多いと考えられる．特に歌声には他の楽器にはない性質があり，主なものを以下に 3 つ挙げる．1 つめはピッチの揺らぎである．短時間の区間で観測した場合，楽器音は一定の音高のピッチを保つ傾向にあるのに対し，ボーカル音声は常にピッチが変化する傾向にある．橘らはボーカル音声のこの性質をメロディ推定に用いているが [9]，2 パスの処理は複雑で処理量が大きく，家電などの組み込み処理系には不向きであった．2 つめは例えばステレオ音源であれば定位位置（パン）がある．同じ周波数帯の楽器音と区別してメロディを際立たせるため，ボーカル音声は他の楽器と異なる定位位置にミックスされることが多いことが考えられる．3 つめが連続性である．楽器の中には数オクターブもの差の 2 つの音を交互に演奏できるものもあるが，ボーカルの音声はピッチが急激に変化することは少ない．

以上より，我々はボーカル音声の 3 つの性質を利用する．
性質 1：ボーカル音声のピッチは常に変化する．
性質 2：他の楽器音と区別してミックスされる．
性質 3：ボーカル音声のピッチはなめらかに変化する．

全体のフロー図を図 1 に示す．まず，性質 1 および 2 を用いてボーカル成分を強調する．次に，性質 3 を用いてボーカル強調されたスペクトル特徴量からピッチ軌跡とピッチトレンドを抽出する．最後に，新しく定式化された動的計画法を用いて最適なメロディラインを決定する．

2.2 ボーカル成分の強調

2.2.1 ピッチ変化による強調

前述した性質 1 である恒常的な音声ピッチの変化は，人間の声の自然な性質である．ボーカル音声は楽器音のように一定の周波数を厳密に保つことは難しい．このようなボーカル音声と楽器音の違いは解析フレーム長を長くした場合に特に顕著になる．パワースペクトルの楽器音由来のピークは急峻となるのに対し，ボーカル音声のそれはなだらかとなる．パワースペクトルの一例（図 2-(a)）を見てみると，ギター由来のピーク（点線の矢印）は鋭く，ボーカル音声由来のものは（実線の矢印）はなだらかである．これらを区別するため，パワースペクトルに対し周波数軸方向にローパスフィルタを処理することで楽器音由来のピークが抑圧されることが期待できる．図 2-(a) のパワースペクトルにローパスフィルタを適用したものは図 2-(b) のようになり，急峻なピークが抑圧されているのが分かる．

このボーカル音声強調を明示的に利用するため，0 から 1 の範囲で以下のように正規化を行う．離散的な時刻インデックス x，周波数インデックス y のパワースペクトルの値を $Y(x,y)$ とする．ただし $1 \leq x \leq X$，$1 \leq y \leq Y$ とする．ローパスフィルタの係数が $\{a_0, a_1, \ldots, a_{K-1}\}$ であるとき，対数パワースペクトルへ畳み込んだ値は

$$l(x,y) = \sum_{k=0}^{K-1} a_k \cdot \log|Y(x,y)| \tag{1}$$

となり，スペクトル特徴量は以下のように正規化することができる．

$$v(x,y) = \begin{cases} 1 & (\mu(x) < U_Y(x,y) < l(x,y)) \\ 0 & (U_Y(x,y) \leq \mu(x)) \\ \dfrac{l(x,y) - \mu(x)}{U_Y(x,y) - \mu(x)} & (otherwise) \end{cases} \tag{2}$$

ただし，$\mu(x)$ は時刻インデックス x の対数パワースペクトル $\log|Y(x,y)|$ における平均値で，$U_Y(x,y)$ は周波数インデックス y から近傍のスペクトルピークのインデックスのうち小さい方を $p_-(y)$，大きい方を $p_+(y)$ として，そのピーク値をそれぞれ $P^-(x,y) = \log|Y(x,p_-(y))|$ と $P^+(x,y) = \log|Y(x,p_+(y))|$ として，

$$U_Y(x,y) = \frac{(p_+(y) - y)P^-(x,y) + (y - p_-(y))P^+(x,y)}{p_+(y) - p_-(y)} \tag{3}$$

のように定義されるピーク同士を直線で結んだ関数．ただしスペクトルピークは以下の条件を満たす全てのインデックス p である．

$$Y(x,p) = \max_{p-H \leq j \leq p+H} Y(x,j) \tag{4}$$

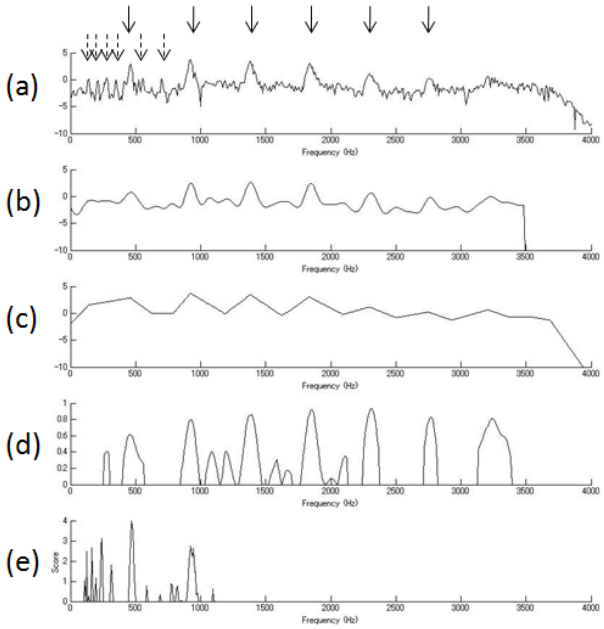

図 2 スペクトル特徴量計算の例: (a) は対数パワースペクトルでボーカル音声由来のゆるやかなピーク（実線の矢印）とギター音由来の急峻なピーク（点線の矢印）を含む．(b) は (a) に対しローパスフィルタを周波数軸方向に畳みこんだ結果．(c) は正規化のための関数 $U(x,y)$．(d) は正規化されたスペクトル特徴量 $v(x,y)$．(e) はボーカルピッチスコア特徴量 $S(x,y)$．

ただし H はパラメータである．

正規化のための関数 $U_Y(x,y)$ は図 2-(c) のようになり，正規化された特徴量 $v(x,y)$ は図 2-(d) のようになる．パワースペクトルのゆるやかなピークのみが $v(x,y)$ では 1 に近い値となっている．

2.2.2 ミキシングによる強調

入力信号が複数楽器をミキシングしたステレオ録音であったとき，性質 2 を用いてボーカル成分を更に強調することができると考えられる．Barry らは 2 チャンネルのスペクトルのパワーを比較することによって，音源の定位位置によって個々の音源を抽出する方法を提案している [12]．また，調波構造の情報を用いることで基本周波数をより効果的に検出する方法は様々な研究で検討されている [7], [11]．これら Barry らの手法と調波構造の情報とを組み合わせてスペクトル特徴量のボーカル成分を強調する．

$A_{i,c}(x,y)$ を

$$A_{i,0}(x,y) = \frac{\sum_{n=1}^{N} v(x,ny)|Y_L(x,ny) - \beta i Y_R(x,ny)|}{N^{\alpha}} \quad (5)$$

$$A_{i,1}(x,y) = \frac{\sum_{n=1}^{N} v(x,ny)|Y_R(x,ny) - \beta i Y_L(x,ny)|}{N^{\alpha}} \quad (6)$$

と定義する．ただし N は解析周波数区間に存在する倍音数，α はパラメータであり，$Y_L(x,y)$ と $Y_R(x,y)$ はそれぞれ左チャネルと右チャネルのパワースペクトルである．また，離散的な定数 β_i は $0 \leq \beta_i \leq 1$ を満たす．このとき，ミキシング情報によってボーカル成分が強調されたスペクトル特徴量を

$$S(x,y) = \max_{i,c} A_{i,c}(x,y) - \min_{i,c} A_{i,c}(x,y) \quad (7)$$

とする．

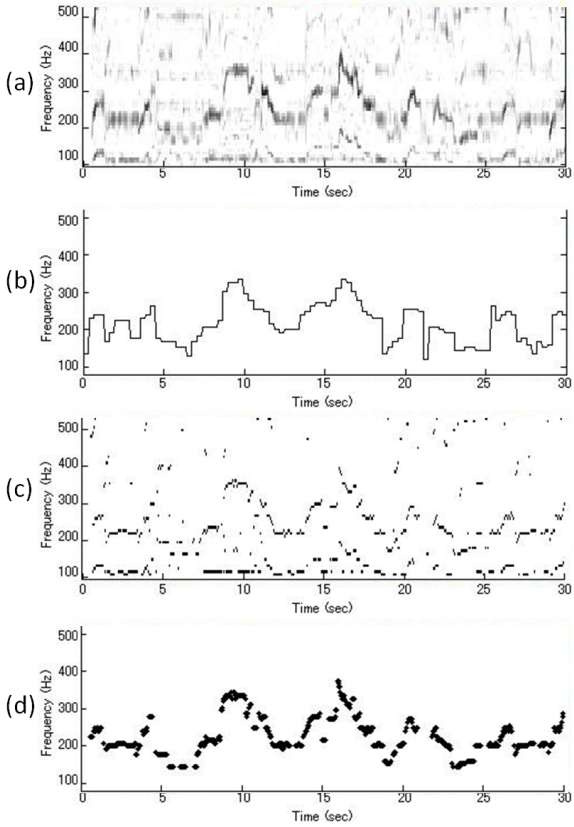

図 3 The Beatles の「Hey Jude」のボーカルピッチスコア $S(x,y)$ からのピッチ推定の例: (a) はボーカルピッチスコア $S(x,y)$，(b) はピッチトレンド $T(x)$，(c) ボーカルピッチスコアから抽出したピッチ軌跡ユニット $S(x,y)$，(d) は動的計画法で選択されたメロディライン．

図 2 のパワースペクトルに適用すると，上記のスペクトル特徴量 $S(x,y)$ は図 2-(e) のようになり，ボーカル成分が強調された同じ定位位置の全ての倍音成分が加算されることになる．図から分かるように，ボーカル音声の基本周波数が強調されている．今後この特徴量 $S(x,y)$ をボーカルピッチスコアと呼ぶことにする．The Beatles の曲「Hey Jude」におけるボーカルピッチスコア $S(x,y)$ の例は図 3-(a) のようになる．

入力信号がモノラルの場合 $Y_R(x,y)$ の値は 0 となり，スコアは

$$S'(x,y) = \frac{1}{N^{\alpha}} \sum_{n=1}^{N} v(x,ny)|Y_L(x,ny)| \quad (8)$$

となる．

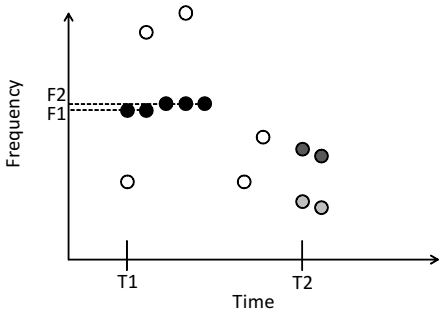

図 4 ピッチ軌跡ユニットの抽出の例．丸はボーカルピッチスコアのピーク位置を表し，色で塗られた丸のまとまりはピッチ軌跡ユニットである．

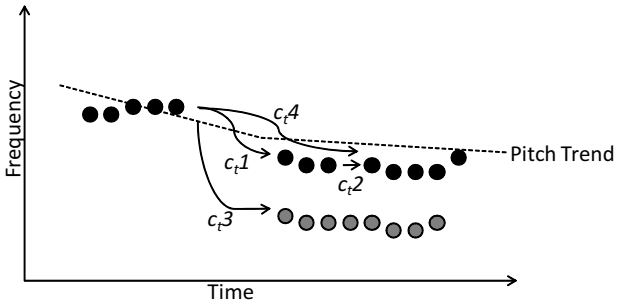

図 5 動的計画法によるピッチ軌跡ユニットの選択の例．c_t1 から c_t4 は遷移コストを表す．

2.3 ピッチ軌跡ユニットの抽出

ボーカルピッチスコア $S(x,y)$ が計算されたら，性質 3 に基づいてピッチ軌跡ユニットを抽出する．我々はピッチ軌跡ユニットをメロディラインの一部となりうる連続したピッチ系列のまとまりと定義する．ボーカルピッチスコアの値が大きいほどメロディラインとなりうる可能性が高いと考えられるため，ピッチ軌跡ユニットはボーカルピッチスコアのピークから抽出する．ピークの抽出は式 (4) と同様に行われる．スコア $S(x,y)$ のピークの中から連続的なものをひとまとまりにし，ピッチ軌跡ユニットとする．連続的とは例えば，隣り合う解析フレーム同士のピーク周波数の差が半音以内である，などと定義できる．図 4 にこの処理を図示する．それぞれの丸がボーカルピッチスコア $S(x,y)$ のピーク位置とする．時刻 $T1$ のとき周波数 $F1$ にピークがあり，続くフレームで周波数 $F2$ にピークが続く場合，もし $2^{-1/12} \leq F2/F1 \leq 2^{1/12}$ を満たす場合これらのピークは黒丸のように 5 フレームでひとまとまりとなる．同様に時刻 $T2$ からのピークもピッチ軌跡ユニットとしてまとめられる．図 3-(a) から抽出されたピッチ軌跡ユニットは図 3-(c) のようになる．

2.4 動的計画法によるピッチ軌跡ユニットの選択

ここで，ピッチ軌跡ユニットからメロディラインを選択する際，再び性質 3 を用いてピッチ軌跡ユニット間の連続性が利用できると考えられる．また，あらゆる組み合わせの中から最適化を行うために動的計画法による最適化が有効であると考えられる．すでにいくつかの研究でメロディラインを選択する際，動的計画法や隠れマルコフモデルのための Viterbi サーチが用いられている [8], [10]．しかし，楽曲全体のメロディラインの連続性とピッチ軌跡ユニット間の連続性のどちらも考慮するためには，動的計画法を新しく定式化し，複雑なピッチ軌跡ユニット同士の組み合わせの中から最適なものを選択する必要がある．全体的な連続性を考慮するためには [8] で提案されているピッチトレンドを利用することが考えられる．ピッチトレンド $T(x)$ はボーカルピッチスコア $S(x,y)$ の時間解像度および周波数解像度を低くすることで [8] のように計算することが考えられ，図 3-(a) のボーカルピッチスコアから計算されたピッチトレンドは図 3-(b) のようになる．

時刻インデックスと周波数インデックスを用いて $\mathbf{U}_m = \{(x_{m,1}, y_{m,1}), (x_{m,2}, y_{m,2}), ...\}$ をピッチ軌跡ユニットとする．ただし，$x_{m,1} = x_{m,2} - 1 = x_{m,3} - 2...$ であり，m はピッチ軌跡ユニットのインデックスである．$\mathbf{U}_m$ の全ての組み合わせから動的計画法を用いて最適なメロディラインを選択する．m の集合を M とし，$\mathbf{U}_m$ の最初のインデックスを $x_{m,\mathrm{first}} = x_{m,1}$ として最後のインデックスを $x_{m,\mathrm{last}}$ とし，$x_{m,\mathrm{last}} < x_{n,\mathrm{first}}, (m < n,\ m,n \in \mathbf{M})$ とする．以下に定義する $D_{\mathbf{M}}$ を最大化することで，ピッチ軌跡ユニットの最適なコンビネーションを決定する．

$$\begin{aligned} D_{\mathbf{M}} = \quad & \textstyle\sum_{m\in\mathbf{M}}\big(\sum_{(x,y)\in\mathbf{U}_m} S(x,y) \\ & -\gamma_1 \textstyle\sum_{(x,y)\in\mathbf{U}_m} |\log y - \log T(x)| \\ & -\gamma_2 |\log y_{m',last} - \log y_{m,first}|\big) \end{aligned} \tag{9}$$

ただし，m' は m 番目のピッチ軌跡ユニットの直前のインデックスで，γ_1 と γ_2 はパラメータである．式 (9) の第 3 項は遷移コストを表す．図 5 にこの説明を図示する．この例では 4 つの遷移コストが図示されており（c_t1 から c_t4），黒色のピッチ軌跡ユニットが集合 $\mathbf{M}$ として選択されている場合，c_t1 と c_t2 のみが考慮される．

この動的計画法は以下の要領で実装される．2 種類の行列を用意し，累積スコアを格納する．$\mathbf{D}_{\mathrm{L}}$ を $Y \times X$ の行列とし，局所的なスコアを格納し，その要素を $d_{\mathrm{L}}(x,y)$ とする．また，$\mathbf{D}_{\mathrm{E}}$ を $Y \times 1$ の行列とし，ピッチ軌跡ユニット終点のスコアを格納し，その要素を $d_{\mathrm{E}}(y)$ とする．まずは終点スコア行列を

$$d_{\mathrm{E}}(0) = 0 \tag{10}$$

とし，局所スコア行列を

$$d_{\mathrm{L}}(0,y) = S(0,y) + \gamma_1 |\log y - \log T(0)| \tag{11}$$

として初期化する．更新規則は

if $(x.y) \in \mathbf{U}_m$ *and* $x = x_{m,\mathrm{first}}$ *then*

$$d_{\mathrm{L}}(x,y) \leftarrow \max_j \left\{ \begin{array}{l} d_{\mathrm{E}}(j) + S(x,y) - \gamma_1 |\log y - \log T(x)| \\ -\gamma_2 |\log j - \log y| \end{array} \right\}$$

(12)

$$if\ (x,y) \in \mathbf{U}_m\ and\ x = x_{m,\mathrm{last}}\ then$$

$$d_{\mathrm{E}}(y) \leftarrow \max \begin{cases} d_{\mathrm{E}}(y) \\ d_{\mathrm{L}}(x-1,y) + S(x,y) - \gamma 1|\log y - \log T(x)| \end{cases} \tag{13}$$

$$if\ (x,y) \in \mathbf{U}_m\ and\ x_{m,\mathrm{first}} < x < x_{m,\mathrm{last}}\ then$$

$$d_{\mathrm{L}}(x_{m,n}, y_{m,n}) \leftarrow \begin{array}{l} d_{\mathrm{L}}(x_{m,n-1}, y_{m,n-1}) + S(x_{m,n}, y_{m,n}) \\ -\gamma_1|\log(y_{m,n}) - \log T(x_{m,n})| \end{array} \tag{14}$$

となり，$D_{\mathbf{M}}$ の最大値は

$$D_{\mathbf{M}_{\max}} = \max_j d_{\mathrm{E}}(j), \tag{15}$$

であり，$D_{\mathbf{M}_{\max}}$ をバックトレースすることによってボーカルメロディーラインが推定される．

図 3-(c) のピッチ軌跡ユニットに対して，動的計画法によって選択されたメロディラインが図 3-(d) となる．

3. 鼻歌検索システムの実装

推定されたボーカルメロディ情報を鼻歌検索に用いる場合，Zhu らの研究に基づいてシンプルな動的計画法が利用できると考えられる [2]．[2] では処理速度を高速化するためのインデキシングについても言及されているが，ボーカルメロディ推定の評価のためであればその部分を実装する必要性はないと考えられる．

事前に提案手法によってボーカルメロディが推定され，検索時にはモノラルのマイク信号による入力クエリに対しピッチ推定が行われる．このピッチ推定は本論文で提案しているボーカルメロディ推定と同様の処理で行うことができる．事前に推定された参照ボーカルメロディ情報と検索時に推定された入力クエリのピッチ情報に対してあらゆる参照ボーカルメロディ情報の切り出しに対して上記動的計画法でロバストにマッチングが行われる．対象楽曲は類似度に応じてソートすることができ，例えば上位 10 曲のみを結果として表示するなどの方法が考えられる．

4. 評価

本論文で提案するボーカルメロディ推定手法に関してメロディ推定の性能と鼻歌検索の性能の 2 つの視点から評価を行った．全ての評価において，ハニング窓を用いた短時間フーリエ変換を用いており，フレームサイズは 2048 点，512 点のシフトで行った．また，入力楽曲の音響信号は 8000Hz のサンプリングレートである．パラメータは実験的に以下のように決定した．ピーク検出時の式 (4) 中のパラメータ H は 16，式 (9) 中のパラメータ γ_1 と γ_2 はそれぞれ 0.52 と 5.2，式 (5) 中の α は 0.5 とし，ローパスフィルタ a_k のカットオフ周波数は 0.15π ラジアンとした．

表 1 ADC2004 データセット（モノラル）における 1/2 半音以内の誤差で推定されたフレームの割合

	Pop	Daisy	Opera	Jazz	MIDI
Proposed	**0.726**	**0.942**	0.549	0.741	0.556
MIDI	0.692	0.907	**0.647**	**0.862**	**0.661**

表 2 ADC2004 データセット（モノラル）における 1 半音以内の誤差で推定されたフレームの割合

	Pop	Daisy	Opera	Jazz	MIDI
Proposed	**0.791**	**0.971**	**0.695**	0.767	0.564
MIDI	0.728	0.911	0.673	**0.874**	**0.675**

4.1 ボーカルメロディ推定の性能評価

まずはボーカルメロディの推定性能について評価を行った．メロディ推定は世界中で広く研究されており，年に 1 度 MIREX[*1] で性能比較が行われている．そこでは共通のデータセットを用い，評価基準としてメロディ音の存在判定，ピッチの正確さなどがあるが，鼻歌検索用途においてはメロディ音の存在しない部分もピッチ推定しておくことで，推定されたピッチ全てと比較を行うことができるため，ピッチ推定の正確さのみで評価すればよい．本論文では MIREX の評価基準，Raw Pitch Accuracy に倣って，ラベル付けされた正解ピッチと 1/2 半音以内の誤差（50 セント）で推定されたフレーム数（フレーム毎のメロディ音の存在の推定エラーを無視する）の，メロディ音が存在するとラベル付された全フレーム数に対する割合として評価する．また，用いたデータセットは MIREX に用いられる ADC2004 とよばれる，5 ジャンル 20 曲の正解ピッチラベルが付与されたモノラル音源のデータセットを用いた．ジャンルは Pop，Daisy，Opera，Jazz, MIDI の 5 つである．また，MIREX でも優れた性能を示した MELODIA[11] との比較を行った．

ボーカルメロディ推定の性能評価の結果を**表 1** に示す．鼻歌検索に用いるためのボーカル音声に特化した手法であるため，メロディ音がボーカル音声でなく楽器音である Jazz および MIDI は性能が十分ではないが，ボーカル音声を含む Pop や Daisy では十分な性能が得られた．Opera は性能として不十分のように見えるが，これはビブラートが非常に大きく，提案手法での時間周波数解像度では追従できていないことに起因していると考えられる．特に鼻歌検索においては人によって歌われた入力音声自体にピッチの不正確性を持っているため，必ずしも 1/2 半音以内の誤差で推定する必要はなく，例えば 1 半音の誤差以内で同様の評価を行った場合，**表 2** のようになり，Opera でも十分な性能を示していることが見てとれる．

[*1] Music Information Retrieval Evaluation eXchange (MIREX) [Online]. http:www.music-ir.org/mirex/wiki/Audio_Melody_Extraction

表 3　MIDI をクエリとした場合の
ポピュラー音楽 1000 曲を対象とした鼻歌検索の性能

Reference Melody	Top 1	Top 5	Top 10
MIDI	100 %	100 %	100 %
MELODIA	48.78 %	51.22 %	51.22 %
Proposed (Monaural)	64.63 %	75.61 %	78.05 %
Proposed (Stereo)	80.49 %	85.37 %	89.02 %

表 4　14 名の鼻歌・歌唱をクエリとした場合の
ポピュラー音楽 1000 曲を対象とした鼻歌検索の性能

Reference Melody	Top 1	Top 5	Top 10
MIDI	78.99 %	86.23 %	89.86 %
MELODIA	47.83 %	50.72 %	53.62 %
Proposed (Monaural)	65.85 %	73.17 %	76.83 %
Proposed (Stereo)	71.74 %	80.43 %	84.06 %

4.2　鼻歌検索の性能評価

次に，推定した参照ボーカルメロディ情報を用いた場合の鼻歌検索における性能も評価した．まず，入力クエリとして，人手で作成した MIDI 情報を用いて，入力鼻歌の不正確性を排除して評価を行った．我々の提案する手法でモノラル音源およびステレオ音源から推定されたボーカルメロディと MELODIA，MIDI それぞれを参照データとした場合の比較を行った．対象楽曲は邦楽および洋楽からなるポピュラー音楽 1000 曲を対象とし，入力クエリの MIDI は対象楽曲の中からランダムに選択した 82 曲のうち 15 秒間分を用いた．**表 3** に結果を示す．Top 1 はクエリと最も類似した楽曲として正解の楽曲が選ばれたもの，Top 5 はクエリと類似する上位 5 曲に正解楽曲が存在したもの，Top 10 は上位 10 曲に正解楽曲が存在したクエリの割合を表す．モノラル楽曲で比較しても我々の提案する手法を用いて推定したボーカルメロディ情報を参照データとした場合の方が優れた性能を示し，ステレオ情報を利用した場合は更に精度よく検索できていることが分かる．

次に，実際被験者に歌ってもらったクエリを用いて鼻歌検索における性能の評価を行った．被験者は 14 名のアマチュア男女であり，上記と同じ対象楽曲 1000 曲を用い，その中の楽曲の中からランダムに 15 秒間の 138 クエリを鼻歌および歌唱で録音した．上記評価と同様にモノラル音源およびステレオ音源から推定されたボーカルメロディと MELODIA，MIDI それぞれを参照データとした場合の比較を行った．**表 4** に結果を示す．当然，MIDI を参照データとして検索を行った場合の性能は一番良いが，我々の提案するボーカルメロディ推定手法のうち特にステレオ情報を利用した場合は十分な性能が示された．

5.　まとめ

本論文では鼻歌検索の参照メロディーデータとして用いることを主目的にしたボーカルメロディ推定手法について述べた．検索対象として最もありうるであろうボーカル音声を以下の 3 ステップで強調することで推定を行った．まず，パワースペクトルに対し周波数軸方向にローパスフィルタを処理して楽器音成分を強調し，定位位置毎の倍音成分重畳処理を行い，ピッチトレンドとピッチ軌跡ユニット抽出による連続性を考慮した動的計画法を用いて最適なボーカルメロディライン推定を行った．実験では特にボーカル音声を含んだポピュラー音楽において優れたボーカルメロディ推定性能と，推定されたボーカルメロディを参照データとした鼻歌検索における高い性能を示した．

今後の展望として，メロディが存在するかしないかの判定を行うことで，メロディ推定自体の性能を向上させることや，ボーカル音声以外のメロディ音への対応が考えられる．また，鼻歌検索の性能向上のため，リズムの検出や歌詞の認識との組み合わせによるラップ音楽への対応などがある．最終的にはビート検出やコード認識との組み合わせによって自動採譜にも応用できると考えている．

参考文献

[1] S. Pauws, “CubyHum: a fully operational query by humming system,” in *Proc. ISMIR*, pp. 187–196, 2002.
[2] Y. Zhu and D. Shasha, “Warping indexes with envelope transform for query by humming,” in *ACM SIGMOD Int. Conf.*, pp. 181–192, 2003.
[3] A. Ghias, J. Logan, D. Chamberlin, and B. Smith, “Query by humming: musical information retrieval in an audio database,” in *Proc. ACM Multimedia*, pp. 231–236, 1995.
[4] L. Lu, H. You, and H.-J. Zhang, “A new approach to query by humming in music retrieval,” in *Proc. ICME*, pp.22-25, 2001.
[5] T. Nishimura, H. Hashiguchi, J. Takita, J. X. Zhang, M. Goto and R. Oka, “Music signal spotting retrieval by a humming query using start frame feature dependent continuous dynamic programming,” in *Proc. ISMIR*, pp. 211-218, 2001.
[6] J. Song, S. Y. Bae, and K. Yoon, “Mid-level music melody representation of polyphonic audio for query-by-humming system,” in *Proc. ISMIR*, pp. 133–139, 2002.
[7] M. Goto, “A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass line in real-world audio signals,” in *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.
[8] Y. Li, D. L. Wang, “Separation of singing voice from music accompaniment for monaural recordings,” in *Trans. Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1475–1487, 2007.
[9] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, “Melody line estimation in homophonic music audio signals based on temporal-variablity of melody source,” in *Proc. ICASSP*, pp. 425–428, 2010.
[10] C. L. Hsu, D. L. Wang, and J.-S. R. Jang, “A trend estimation algorithm for singing pitch detection in musical recordings,” in *Proc. ICASSP*, pp. 393–396, 2011.
[11] J. Salamon and E. Goméz, “Melody extraction from polyphonic music signals using pitch contour characteristic,” in *Trans. Audio, Speech and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.
[12] D. Barry, B. Lawlor, and E. Coyle, “Sound source separation: Azimuth discrimination and resynthesis,” in *Proc. Int. Conf. Digital Audio Effects*, 2004.
[13] R. B. Dannenberg and N. Hu, “Understanding search performance in query-by-humming systems,” in *Proc. ISMIR*, pp.232–237, 2004.