Vol.2015-MUS-107 No.56
2015/5/24

フレーム間の音程に注目した歌唱音声からのメロディ抽出法

鈴木 基之[1,a)]　久保 勇人[,†1]

概要：本報告では，歌声を用いた音楽検索システムに用いるため，歌唱音声からメロディ（音程系列）を高精度に抽出する方法（MEMTIS 法）を提案する。2 つのフレームから対数周波数スペクトルをそれぞれ求め，その相互相関関数の最大値をとることで，音程を計算することができる。子音部等，不要なフレームを除去した上ですべてのフレームの組み合わせについてこの計算を行い，そこから誤差最小となる音程系列を求める。歌唱音声に対してメロディの抽出実験を行ったところ，従来法による 1 フレームあたりの抽出誤差 73.1 cent に対して，提案法は 50.1 cent であり，有効性が示された。

1. はじめに

近年音楽圧縮技術が向上し，大量の音楽データを mp3 プレーヤーなどの小型デバイスに保存，再生することが可能となった。これらのデバイスは小型であるがゆえに，一般的にインターフェイスが貧弱である。その一方で保存可能な曲数は何千曲と膨大な数になっているため，保存してある曲の中から特定の 1 曲を指定して再生する，といった操作が非常にやり難い仕様となっている。

こうした問題を解決するため，歌声やハミングを用いて楽曲を検索するシステム（Query-by-Singing/Humming）が開発されてきた。こうしたシステムにおいては，データベース中の楽曲（通常は MIDI といった楽譜データで表現される）がそのまま入力されるわけではないことから，音符の情報（音の高さや長さ）を抽出し，それらを用いて検索を行う必要がある。多くのシステム（例えば，[1–3] 等）では，こうした特徴量の抽出を容易にするため，入力歌唱を「タタタ」等のハミングによる歌唱に限定している。こうすることで，音符の区切り時刻は入力音声のパワー情報から推定し，また音符の高さ情報は「タ」に含まれる母音の安定部からピッチを抽出する方式を採用している。

しかし一方で，こうした歌唱方法は自然ではないため，歌詞による歌唱音声を入力としたい，という要望も多い。そこでいくつかのシステムでは，歌詞による歌唱を入力可能とするよう対処が行われているが，こうしたシステムにおいても，音の高さの抽出は従来からのピッチ抽出法が用いられている。一般にピッチ抽出法の多くは，一定の波形が安定して繰り返している区間（母音の安定部等）ではよい推定が行えるが，そうではなく波形が乱れている区間（子音部やそれらとの結合区間等）では，精度よく推定することができない。なんらかの方法で音符の区切り時刻が推定できている場合は，その情報を元に母音部に相当するフレームのみピッチを計算し，その値をもって音符の高さ情報にすることが可能であるが，音符の区切り時刻を推定すること自体が難しく，逆にピッチ情報等を用いて区切り時刻を推定する方法も提案されている [4–6] ほどであり，鶏と卵の関係とも言える。

音の高さの推定法では，楽曲検索においては「高さ」そのものの情報ではなく，相対的な高さ（隣りの音符に対する高さの差）がわかれば十分であるため，2 つのフレーム間の相対的な高さ（音程）を推定する方法 [7] が提案されている。この方法を用いれば，ピッチ抽出法を用いるよりも頑健に音程を推定することが可能である。しかし，この方法でも波形が不安定な部分の推定を高精度に行うことはできない。

そこで本報告では，事前に推定に用いるには不適切であるフレームを排除し，また複数のフレームを用いて推定値を求め，それらから誤差最小となる系列を自動推定することで，より頑健に音程系列を推定する方法を提案する。

2. フレーム間の音程に注目したメロディ情報の抽出

2.1 相互相関関数を用いた音程抽出法の概要

市川らの方法 [7] は，相互相関関数を用いて 2 つのフレーム間の音程を高精度に抽出する。ここでは，その方法の概要を説明する。

[1] 大阪工業大学
Osaka Institute of Technology, 1-79-1, Hirakata, Osaka, 573-0196, Japan
[†1] 現在，lAVC マルチメディアソフト（株）
Presently with AVC Multimedia Software Co., Ltd.
[a)] moto@m.ieice.org

ⓒ 2015 Information Processing Society of Japan

一般に有声音のスペクトルは，基本周波数に加えて 2 倍音，3 倍音等が含まれる，いわゆる調波構造を持っている。例えば基本周波数が 200Hz の有声声であれば，400Hz，600Hz，800Hz といった周波数にも成分を持つようなスペクトルとなっている。

ここで，基本周波数 f_a[Hz] を持つ有声声の n 倍音の周波数は nf_a[Hz] である。同様に，基本周波数 f_b[Hz] を持つ有声声の n 倍音の周波数は nf_b[Hz] となる。それぞれ対数をとり，その差 d_{ab} を計算すると，

$$\begin{aligned} d_{ab} &= \log(nf_a) - \log(nf_b) \\ &= \{\log(n) + \log(f_a)\} - \{\log(n) + \log(f_b)\} \\ &= \log\left(\frac{f_a}{f_b}\right) \end{aligned} \quad (1)$$

となり，n にかかわらず一定の値であることがわかる。つまりこの式は，横軸を対数周波数としたスペクトルにおいて，基本周波数の異なるふたつの音の倍音構造は，すべての n 倍音が同じ量 d_{ab} だけ平行移動していることを示しており，その移動量 d_{ab} は両者の基本周波数の比（=音程）であることがわかる。

そこで，ふたつの音のスペクトルの横軸を対数に変換し，相互相関関数が最大値を示す「ずれ量」を計算することで，2 フレーム間の音程を直接計算する方法が，市川らの方法である（図 1）。

この方法では，（周波数軸を対数に変換すると）スペクトル全体の形状がほぼ同じであることを仮定している。そのため，母音と無声子音，といったように調波構造がよく表れているスペクトルと，そうではないスペクトルの間の音程を計算しようとすると，スペクトル形状が大きく異なるために正しく計算できない，という問題点がある。

2.2 MEMTIS 法の概要

あるフレームと別のフレームの間の音程は，前節で説明した市川らの方法で求めることができる。そのため，入力音声の中からある 1 つのフレームを基準フレームとして選択し，そのフレームと他のすべてのフレームとの間の音程を計算すると，入力音声全体の音程系列（メロディ情報）を得ることができる（図 2）。

しかしこの方法では，最初に選択する基準フレームの性質によって推定性能が左右されることが容易に予測できる。例えば無音や無声子音などに対応するフレームを基準フレームとしてしまった場合，正しく音程を計算することができない。更に，慎重に基準フレームを選択したとしても，すべてのフレームとの音程が正しく推定可能である基準フレームが存在するとは限らず，あるフレームは前半のフレームとの音程が推定できず，別のフレームを選択すると，後半のフレームとの音程の推定精度が低下する，といった状況になることも考えられる。

そこで，以下のふたつの対処法を導入する。

- 計算に用いるフレームを事前に選択する
  市川らの方法では主に調波構造に注目して音程を計算することから，調波構造が強く表れていないフレームを事前に抽出し，それらを除いた上で音程系列の計算を行うこととする。除かれたフレームに対応する時刻の音程は，前後の音程系列から線形補間することで求める
- 計算に用いるすべてのフレームを基準フレームとして選択し，その他のフレームとの音程を計算する，ということを繰り返すことで頑健性を向上させる
  すべてのフレーム間の音程が正しく推定できるのであれば，どのフレームを基準としたとしても，すべて同じ音程系列が得られるはずである。しかし，前述したように一部のフレームの組みあわせで音程推定精度が劣化する，といったことが考えられる。そこで，すべてのフレームをそれぞれ基準フレームとして他のフレームとの間の音程系列を求める，ということを繰り返し，それらの結果から誤差が最小となる音程系列を求めることで，推定精度の劣化に左右されにくい頑健な音程系列の推定を行う（図 3）。

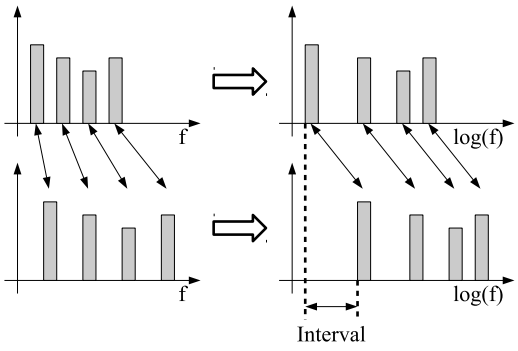


図 1 相互相関関数を用いた音程抽出法の概要

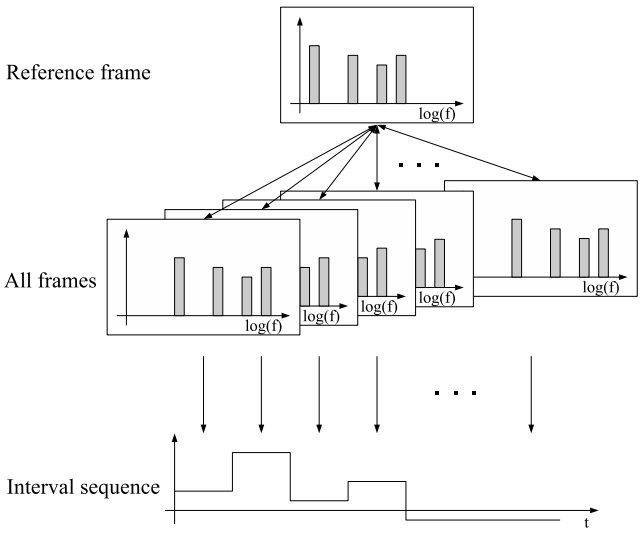


図 2 基準フレームとの間の音程系列の推定

最終的に，提案するアルゴリズム（MEMTIS 法：Melody Extraction Method based on Tone Interval Sequences）は以下のようになる。

(1) パワーを用いて，有声区間を抽出
(2) 有声区間中で調波構造が強く表れているフレームを抽出
(3) 抽出されたすべてのフレームについて，他のフレームとの間の音程を計算
(4) 最小二乗法を用いて，誤差最小基準で音程系列を推定

次節以降では，各ステップについて詳細に説明する。

2.3 MEMTIS 法における各ステップの詳細

2.3.1 有声区間の抽出

入力から実際に歌唱しているフレームを抽出する。これは音声認識システムでいうところの VAD（Voice Activity Detection）に相当する。また，歌唱中であっても息継ぎや休符に相当する区間等，ショートポーズに相当するフレームも存在する。こうしたフレームも取り除き，歌唱音声が存在するフレームのみ抽出する。

VAD の方法は様々なものが提案されている [8] が，ここでは簡単にフレーム内の平均パワーを用いる。この時，固定された閾値以上のパワーを持つフレームのみ抽出，といったことを行うと，録音レベル等に性能が左右されてしまうため，入力フレームすべてを用いてクラスタリングを行い，小さな平均パワーを持つクラスタに属するフレームを除去することとする。

予備実験の結果から，k-means クラスタリングを用いて全体を 3 つにクラスタリング（各クラスタは，無音，通常歌唱，一部の大きな音におおよそ対応していることがわかった）し，一番小さな平均パワーを持つクラスタのみを除去することとした。

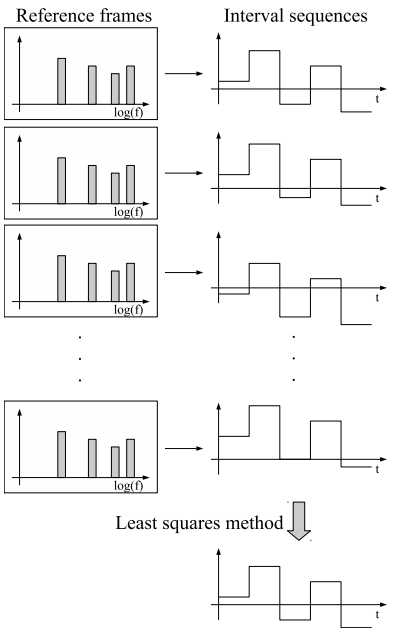


図 3 最小二乗法による音程系列の推定

2.3.2 調波構造が強く表れるフレームの抽出

市川らの方法では，調波構造が音程推定にとってよい手掛りになると思われるため，調波構造が比較的よく表れているフレームを抽出することととした。ここでは，調波構造がよく表れているかどうかの指標として，HNR（Harmonic-to-Noise Ratio）[9] を用いた。

前段で抽出されたすべてのフレームに対し，それぞれ HNR を計算する。その後，HNR が低いフレームを調波構造がよく表れてはいないとして除外する。ここでも有声区間の検出法と同様，固定の閾値を用いることはせず，クラスタリングを用いて除外フレームを決定することとする。

予備実験の結果から，全フレームの HNR の値を 3 混合の混合ガウス分布で近似し，平均値が一番小さなガウス分布との尤度が一番高いフレームを除外フレームとすることとした。有声区間の検出法と同様に k-means 法でも実験を行ってみたが，HNR の場合は混合ガウス分布によるクラスタリングの方がよい性能となったため，こちらを採用した。

2.3.3 音程系列の推定

前述するふたつの方法で抽出されたすべてのフレームを用いて，音程系列の推定を行う。まずは，すべてのフレームの組み合わせについて市川らの方法を用いて音程を推定する。ここで，i 番目のフレームを基準とした時の j 番目のフレームの音程（i 番目のフレームの音に対して，どれだけ j 番目の音が高い/低いか）を $x_i(j)$ と記述する。定義から，$x_i(j) = -x_j(i)$ である。

また，正解となる音程系列を $b(t) = b(1), b(2), \cdots, b(N)$ とする。ここで N はフレーム数である。この時，仮にすべてのフレームでの音程 $x_i(j)$ が正しく計算できているとすると，以下の式が成り立つ。

$$b(t) = x_i(t) + b(i) \tag{2}$$

しかし，実際には推定値 $x_i(j)$ には誤差が含まれている。この誤差の値 $b(t) - \{x_i(t) + b(i)\}$ の総和が最小になるように，最小二乗法を用いて音程系列 $b(t)$ を求める。

二乗誤差の総和 d は

$$d = \sum_i \sum_t \{b(t) - \{x_i(t) + b(i)\}\}^2 \tag{3}$$

となる。この式を $b(1)$ で偏微分すると，

$$\frac{\partial d}{\partial b(1)} = 4\left\{Nb(1) - \sum_i^N b(i) - \sum_i^N x_i(1)\right\} \tag{4}$$

となる。同様に $b(2), b(3), \cdots, b(N)$ でも偏微分し，それぞれ 0 とおくと，以下の連立方程式が得られる。

$$\begin{pmatrix} N-1 & & & -1 \\ & N-1 & & \\ & & \ddots & \\ -1 & & & N-1 \end{pmatrix} \begin{pmatrix} b(1) \\ b(2) \\ \vdots \\ b(N) \end{pmatrix} = \begin{pmatrix} X(1) \\ X(2) \\ \vdots \\ X(N) \end{pmatrix} \tag{5}$$

ⓒ 2015 Information Processing Society of Japan

表 1　実験条件
Table 1　Experimental conditions

歌唱データベース	MIR-QBSH	
	歌唱者	195 名
	データ数	4431 データ
音響分析条件	サンプリング周波数	8 kHz
	量子化ビット数	8 bit
	フレーム幅	64 ms ハニング窓
	フレーム周期	16 ms
評価基準	1 フレームあたりの抽出誤差	

ここで，$X(t) = \sum_{i}^{N} x_i(t)$ である。これを解いて，$b(t)$ を求める。

3. 音程系列の推定実験

提案する MEMTIS 法の有効性を確認するため，歌唱音声からの音程系列の推定実験を行った。

3.1 実験条件

実験条件を表 1 に示す。歌唱データは，国際的な音楽情報処理技術に関するコンテストである MIREX（Music Information Retrieval Evaluation eXchange）[10] の歌唱による楽曲検索テストで利用されている，MIR-QBSH コーパス [11] を利用した。本コーパスには 195 人の歌唱者による 4431 歌唱データが含まれており，そのすべてについてピッチ系列が人手によって付与されている。歌唱内容は，歌詞や鼻歌で様々な曲をアカペラで歌唱しており，一部のデータには収録時の電気的なノイズが含まれている。

ふたつのフレーム間の音程を推定する際には，調波構造がよく表れている帯域のみを使用するため，65Hz から 4kHz までのスペクトルのみを用いた。

また比較対象として，OpenSMILE [12, 13] を用いてケプストラムからピッチを求める方法と，TANDEM-STRAIGHT [14] による結果も計算した。この際，OpenSMILE においては，ピッチの探索範囲（推定する最高周波数）を 500Hz，670Hz のふた通りに設定し，それぞれ結果を比較した。

3.2 音程の抽出精度

各方法における音程の抽出誤差を表 2 に示す。ここで，推定しているのは音程系列であることから，推定系列と正解系列の平均値が一致するように推定系列にバイアスを加算し，その上で各フレームごとの誤差の平均値を計算した。また，正解系列，推定系列ともに無音区間等，音程を抽出できないフレームが存在することから，誤差の計算には両者で音程が抽出できているフレームのみを用い，同時に正解系列で音程が存在するフレーム数に対する割合も「抽出フレーム率」として示した。また「抽出誤差」欄中，スラッシュの前の数字は誤差の絶対値の平均を，後ろの数字は二乗誤差の平均値の平方根（いわゆる RMSE: Root Mean Squared Error）を示している。

表 2　音程の抽出精度
Table 2　Estimation accuracy

抽出法	抽出誤差（cent）	抽出フレーム率
VAD のみ	70.1 / 170.9	86.7%
VAD + HNR	50.1 / 122.4	86.7%
ピッチ抽出（500Hz まで）	75.9 / 173.5	77.8%
ピッチ抽出（670Hz まで）	137.4 / 290.6	77.8%
TANDEM-STRAIGHT	73.1 / 214.5	85.7%

この結果を見ると，MEMTIS 法が最もよい精度を示していることがわかる。平均誤差は 50.1cent（半音の半分程度）であり，検索として用いることを考えた場合，半音の違いも正しく区別できることがわかる。HNR を用いてフレームの選択を行わないと，平均誤差は 70cent 程度まで大きくなることから，HNR によるフレーム選択が想定どおり効果的であったことがわかる。なお今回のデータにおいては，14.4%のフレームが HNR によって除外されていた。

一方，従来からよく用いられているピッチ抽出法では，探索範囲を 500Hz までに制限した場合 75.9 cent，TANDEM-STRAIGHT でも 73.1 cent と，誤りが多いことがわかる。今回使用したデータにはピッチが 500Hz を越えるような歌唱も含まれていたが，ここでは探索範囲を 500Hz までに制限したため，こうした歌唱に対しては必ず誤ったピッチを推定してしまう。そこでこうしたデータに対しても正しくピッチが推定できるよう，探索範囲を 670Hz まで拡げて実験を行ったところ，平均誤差は 137.4cent と，かえって悪化してしまった。これは，ピッチが 500Hz を越えるような歌唱データにも対応できた一方で，低いピッチの歌唱に対して，2 倍音をピッチとして推定してしまった結果，悪化したものと思われる。

そこで，各フレームの誤差の頻度分布を計算した。一部を拡大したものを図 4〜図 6 に示す。これを見ると，ピッチ抽出を行った場合は $\pm 1,200$cent 付近に小さなピークが見られることがわかった。これは，ピッチ抽出時に 2 倍の周波数や半分の周波数に誤抽出してしまったため，およそ 1 オクターブの誤差がでたフレームが一定数あったためと思われる。一方 MEMTIS 法ではそのようなピークは見られず，そうした誤りは少なかったものと思われる。

また，無音に近いフレーム等ではピッチが得られない場合があるが，正解の音程がついているフレーム中で，音程推定結果が得られたフレームの割合は MEMTIS 法では 86.7%であった。これは，最初の VAD において歌唱区間であるにもかかわらず，13%程度のフレームを削除してしまった事を表している。今後はより精度の高い VAD 法の導入を検討する必要があると思われる。一方，ピッチ抽出

ⓒ 2015 Information Processing Society of Japan

による方法では 77.8%と，より低い値となっていた。これは，無声子音等のフレームについてピッチ抽出を行わなかったため（MEMTIS 法は HNR によって削除されたフレームは，線形補間で音程を決定している）と思われる。こうした点からも，MEMTIS 法の方が実用上優れていることがわかる。

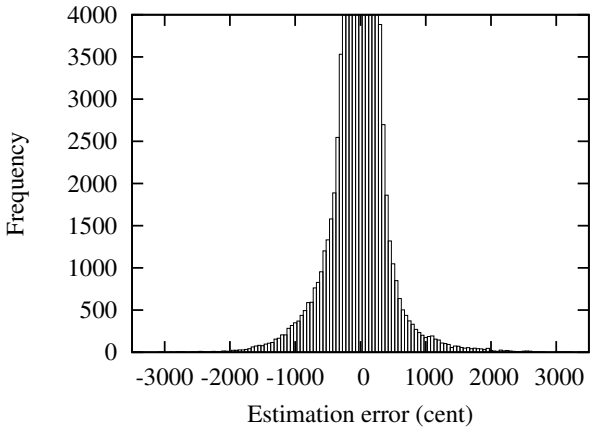


図 4 MEMTIS 法による誤差分布

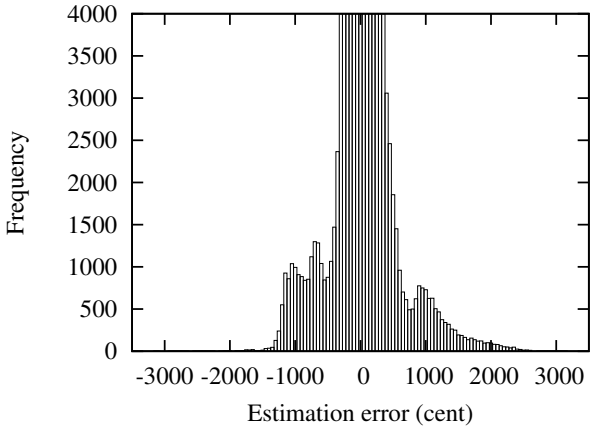


図 5 ピッチ抽出法による誤差分布（500Hz）

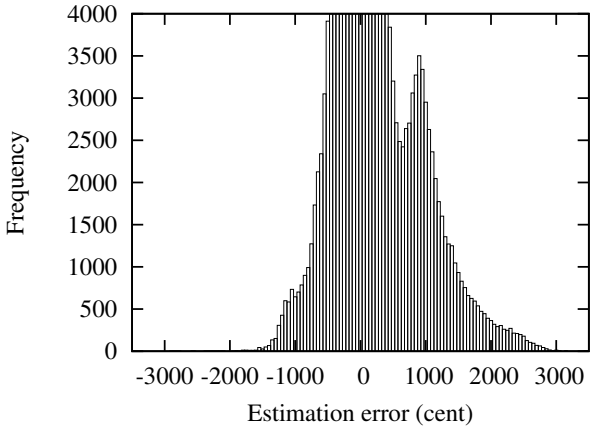


図 6 ピッチ抽出法による誤差分布（670Hz）

4. まとめ

歌詞による歌唱音声を入力とする楽曲検索システムに用いるため，任意の歌唱に対する高精度な音程推定法（MEMTIS 法）を提案した。パワー情報を用いて音声区間の検出を行った後，調波構造が強く表れているフレームを HNR を用いて抽出する。そうして得られたすべてのフレームについて，他のフレームとの間の音程を，対数周波数スペクトルの相互相関関数を用いた音程推定法を用いて推定する。そうして得られた数多くの音程系列を用い，誤差が最小となる音程系列を最小二乗法を用いて推定する。

アカペラによる歌唱音声に対して音程の推定実験を行なったところ，平均で 1 フレームあたり 50cent 程度の誤差で推定できることがわかった。一方，従来からよく用いられている TANDEM-STRAIGHT 法では 73.1cent，ケプストラムによる推定法では 75.9cent の誤差となり，本方法が有効であることが示された。

参考文献

[1] Kosugi, N., Nishihara, Y., Sakata, T., Yamamoto, M. and Kushima, K.: A Practical Query-By-Humming System for a Large Music Database, *ACM Multimedia 2000*, pp. 333–342 (2000).

[2] Ghias, A., Logan, J., Chamberlin, D. and Smith, B. C.: Query By Humming: Musical Information Retrieval in An Audio Database, *Proc. ACM Multimedia*, pp. 231–236 (1995).

[3] Liu, B., Wu, Y. and Li, Y.: A Linear Hidden Markov Model for Music Information Retrieval Based on Humming, *Proc. ICASSP 2003*, Vol. V, pp. 533–536 (2003).

[4] Pollastri, E.: Some Considerations About Processing Singing Voice for Music Retrieval, *Proc. ISMIR*, pp. 285–286 (2002).

[5] Klapuri, A. P., Eronen, A. J. and Astola, J. T.: Analysis of the Meter of Acoustic Musical Signals, *IEEE Trans. Speech, Audio, and Language Processing*, Vol. 14, No. 1, pp. 342–355 (2006).

[6] Birmingham, W., Pardo, B., Meek, C. and Shifrin, J.: The MusArt Music-Retrieval System: An Overview, *D-Lib Magazine*, Vol. 8, No. 2 (2002).

[7] Suzuki, M., Ichikawa, T., Ito, A. and Makino, S.: Novel Tonal Feature and Statistical User Modeling for Query-by-Humming, *IPSJ Journal*, Vol. 50, No. 3, pp. 1100–1110 (2009).

[8] 藤本雅清：音声区間検出の基礎と世界的な研究動向，今後の展開，電子情報通信学会誌，Vol. 95, No. 8, pp. 754–758 (2012).

[9] Boersma, P.: ACCURATE SHORT-TERM ANALYSIS OF THE FUNDAMENTAL FREQUENCY AND THE HARMONICS-TO-NOISE RATIO OF A SAMPLED SOUND, *Proc. Institute of Phonetic Sciences, Univ. Amsterdam*, Vol. 17, pp. 97–110 (1993).

[10] MIREX organizing community: Music Information Retrieval Evaluation eXchange. `http://www.music-ir.org/mirex/wiki/MIREX_HOME`.

[11] Jang, J.-S. R.: QBSH: A Corpus for Designing QBSH (Query by Singing/Humming) Systems. available at

ⓒ 2015 Information Processing Society of Japan

the QBSH corpus for query by singing/humming (http://www.cs.nthu.edu.tw/~jang).

[12] Eyben, F., Wollmer, M. and Schuller, B.: openEAR — Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit, *Proc. 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2009 (ACII 2009)*, Vol. I, pp. 576–581 (2009).

[13] Eyben, F., Wollmer, M. and Schuller, B.: openSMILE — Speech and Music Interpretation by Large-space Extraction. http://opensmile.sourceforget.net/.

[14] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. and Banno, H.: Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation, *Proc. ICASSP 2008*, pp. 3933–3936 (2008).

ⓒ 2015 Information Processing Society of Japan