2005－MUS－62（2）
2005／10／14

周波数領域での調波構造モデルを用いた複数音源のピッチ推定

内田 瑛一　　小坂 直敏

東京電機大学
〒101-8457 東京都千代田区神田錦町 2-2
E-mail: uchida@srl.im.dendai.ac.jp, osaka@im.dendai.ac.jp

あらまし
本報告では音源が複数個混在するモノラル音響信号を対象とし、各音源のピッチ推定を行う手法を提案する。まず、有ピッチ音を対象に、周波数領域において調波構造のモデル化を行う。この調波構造モデルと短時間フーリエ変換による観測スペクトルとの相関に基づく類似度を算出する。得られた相関や類似度の周波数系列において極大値を算出し、これを基に各音源の基本周波数を推定する。最後にピアノの単音を組み合わせたデータに本手法を適用した例を示す。

Pitch estimation of sound mixture using a harmonic structure model in frequency domain

Yoichi UCHIDA, Naotoshi OSAKA

Tokyo Denki University
2－2 Kanda-nishikicho, Chiyoda-ku, Tokyo, 101-8457 Japan
E-mail: uchida@srl.im.dendai.ac.jp, osaka@im.dendai.ac.jp

Abstract
This paper describes a multiple pitch estimation technique for sound mixture of monaural audio signal. First, harmonic structure is modeled in a frequency domain for pitched sound signal. Then correlation based similarity is derived between a harmonic structure model and a observed spectrum acquired from STFT (Short Time Fourier Transform). Local minima are extracted from acquired frequency series of a correlation and a similarity. Pitch of each sound source is estimated based on its local maximum. The technique was applied to mixed piano sounds and its estimation results are shown.

1. はじめに

近年、カラオケ産業や携帯電話の普及に伴う着メロ(着信メロディ)サイトの急増などにより、音楽情報検索、メロディへの自動和声付けなど、音楽情報処理技術分野の重要性が増している。一方、同技術分野の中でも音源分離、楽器音同定、自動採譜などの技術は、音楽家あるいは音楽愛好者にとって、演奏および鑑賞目的の譜面作成のために従来から必要とされてきた。

こうした現状の中で、われわれは音源分離再合成システムの構築を大きな目標として研究を進めている。音楽音響信号から自動的に基本周波数を推定し、譜面を興す自動採譜システム、さらに、音楽信号からメロディのみを独立して再合成できるシステムが達成できれば非常に有用である。

混合音の基本周波数の推定は、それぞれの音源の基本周波数とその倍音成分が重なり、波形やスペクトル形状が複雑になる。その結果、各音源が整数倍の調波構造を有しているにもかかわらず、基本周波数の推定が非常に困難になる。

この問題に対して、これまで様々な研究がなされてきた。くし型フィルタを用いた手法[1]では、従属接続したくし型フィルタを周波数領域で等間隔に零点を配置させ、その零出力を検出することによってピッチの推定を行う。これにより、倍音比を必要とせず零出力を検出するという単純な処理で多声の音楽データのピッチが推定できることを示した。また、柏野らは周波数成分、単音、和音の 3 つの仮説ネットワークを持った階層によって、事後確率最大となる仮説の組を逐次求めていき、これをベイジアンネットワークを用いて情報統合する手法を提案した[2],[3]。この手法では階層構造を構成することによって抽象度の低い順、

もしくは高い順に処理することによって誤認識の軽減を行っている。しかし、これを実現するためには統計データなど大量のテンプレートデータを必要とする。

従来の研究はフレーム単位のアルゴリズムを中心としていた。一方、後藤は最も優勢な基本周波数をEMアルゴリズムを用いて推定する方法を提案した[4]〜[6]。この手法ではマルチエージェントモデルを導入し、時間軸の軌跡も考慮した。また、亀岡らはEMアルゴリズムを拡張して、時間方向まで含めたクラスタリングの問題として定式化した数理モデルを提案した[7],[8]。このモデルは、EMアルゴリズムという枠組みで時間周波数構造としてのスペクトルを扱えるよう拡張し、本問題との親和性が良く、また性能の良い手法として注目されている。

われわれは正弦波モデルによる音源分離再合成を目標としている。ここでの大きな構想は、時間構造が多層になっていることを表現することである。すなわち、フレーム単位からの瞬時ピッチから音符に相当する区間ピッチの検出、またビブラート情報などの時価の異なる情報の統合の問題を扱いたい。また、正弦波モデルを用いた分離再合成のために、分離された一音ストリームの瞬時位相の算出を行いたい。この目的のためにはピッチ推定は前処理である。今回瞬時のピッチ推定では後藤や亀岡らの同研究手法を採用することも考えられたが、時間方向も含めたモデル化に際して、将来的にこれをガウス関数の混合モデルとして扱うよりも減衰、持続、ゆらぎ、変調特性など楽音の特徴を表現するより限定されたモデルを取り込みたい、との立場で別の手法を提案する。

以下で提案する手法は最適化を行うものではないが、相関を用いるという単純な発想で、フレーム単位の最適化よりも階層的な時間構造の統合時、あるいは将来の正弦波モデルへの応用時での親和性を期待している。特に得られる相関関数はミッシングファンダメンタルの補完、調波の抑制など、スペクトルを基本周波数を抽出する視点で是正する、という解釈ができる。このことが、「聴く」という行為は聴覚が音信号を脳内で再構成すること、という捉え方と同等な機能を具備し、正弦波による創造的分離再合成にも繋がる期待がある。

2. 周波数領域での調波構造モデルの検討

周波数領域における調波構造のモデル化にあたり、以下のガウス関数を用いた。

$$g(x,m)=\exp\left(\frac{-(x-m)^2}{2\sigma^2}\right) \qquad (1)$$

ここに、xを確率変数として、mは平均、σ^2は分散を表す。この関数の特徴は単峰性であり、関数も導関数も連続的であることである。これは、周波数領域での基本周波数も含めた一つの調波を表現している。また、この関数は変数が±∞へ行くと関数値が0に漸近する。この特徴は調波の加算を行う際に一つの関数のみが優位になり、他の関数の値はほとんど関与しない、という特性を出すために必要である。

このような関数は他にも$\frac{1}{a+x^2}$なども考えられるが、時間窓関数との整合性の意味でガウス関数を選んだ。また、sinc関数のように、周波数領域でサイドローブが生じ、振動しながら0に漸近する関数は、ローカルピーク抽出時に無駄なピークを作るためにここでは用いない。

(1)式を基にして一つの周波数軸上で調波を表現し、これを周波数軸上で調波に対応させてずらし加算した関数(2)式で表し、これを調波構造のモデルとする。

$$\varphi(\omega,\mu)=\sum_{n=1}^{N}\frac{1}{n}\cdot g(\omega,n\cdot\mu) \qquad (2)$$

ここに、$\varphi(\omega,\mu)$はμを基本周波数としたときの周波数ωでのパワースペクトル値、Nはナイキスト周波数内に入るμの倍音数とする。

現在は音色のモデルは用いていないため、$1/n$の重みは平均的な音色を仮定したものである。

σ^2は時間領域でのガウス関数の窓長によって決まる周波数領域での分散と同じ値を用いた。nを非整数値にして、非調和性の音源に対してもモデル化を行うことが可能だが、今回は調和性を有する音源のみを扱う。

3. 調波構造モデルと観測スペクトルとの相関による基本周波数推定

3.1. 単音の基本周波数推定

ここで観測スペクトルに含まれている基本周波数を強調するため、Ωをナイキスト周波数とし、(3)式によってモデルと観測スペクトルの相関を考える。

$$C(\mu)=\sum_{i=0}^{M} f(\omega_i)\cdot\varphi(\omega_i,\mu) \qquad (3)$$

$$(\Omega=\omega_M\text{とする})$$

これは、ある基本周波数μの(2)式による調波構造モデル$\varphi(\omega,\mu)$と、観測スペクトル$f(\omega)$との相関を表している。そこで、μを順次変化させていき、0[Hz]からナイキスト周波数までの相関値を調べる。

調波構造モデルと、時間領域でガウス窓を用いた音楽信号のスペクトルとの相関を見ることにより、モデルと合致した信号では同一関数となり最

大値となる。これによって基本周波数の整数倍、および整数分の１に強度の高い信号系列が得られる。図 1 にスペクトル、図 2 に(3)式によって得られた相関値を示す。ただし、低い周波数(50Hz 未満)は扱わないものとし、相関値は 0 とした。

単音に関しては、相関によって得られた信号系列に対して、ピーク検出を行うことによって基本周波数を推定できる。有ピッチのスペクトルはミッシングファンダメンタルに代表されるように、必ずしも基本周波数の振幅値が倍音に比べて大きいわけではない。図 3、4 にミッシングファンダメンタルの場合について、観測スペクトルと相関値を示す。これは観測スペクトルで基本周波数の振幅値が倍音成分よりも小さい場合においても基本周波数が検出できる可能性を示している。しかし､現在は最大値として現れているわけではない。

3.2．混合音への適用

以上のことを混合音に拡張する。混合音に適用した場合は、単音で行った結果が複数個加算されると考えられる。そこで混合音に適用した例を図 5、6 に示す。ここでの問題は、音源数が知られていない状態で、どのローカルピークまでを音源の基本周波数として抽出すればよいか、という問題である。しかし、ある基本周波数のピークの振幅値が、違う基本周波数の倍ピッチや半ピッチのピークより小さい可能性があり、閾値を設定することが困難である。しかし、階層的な時間軸の統合を目的としているため、フレーム単位での性能にこだわり過ぎないものとし、相関値が優勢であるローカルピークは抽出されるべきである。このことから、今回は特に音源数の推定を厳しく行わない。

ここで相関を用いた基本周波数と相対的に優勢なローカルピーク抽出までの流れをｉ～ivに示す。

i. 入力された音響信号に対し短時間フーリエ変換を行う
ii. 50Hz～ナイキスト周波数に相当する周波数ビンを基本周波数として調波構造をモデル化
iii. 調波構造モデルと観測スペクトルとの相関を式(３)により計算する
iv. iiiで計算された相関値からローカルピークを抽出する。

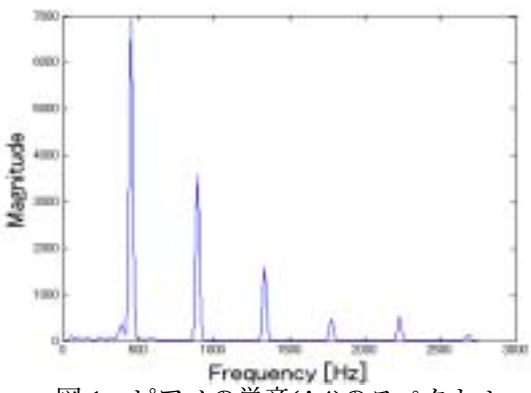

図１　ピアノの単音(A4)のスペクトル

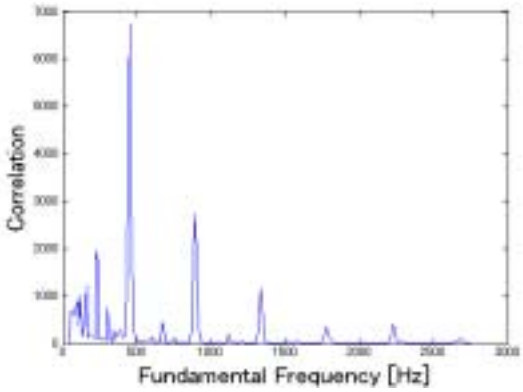

図２　ピアノの単音(A4)の相関

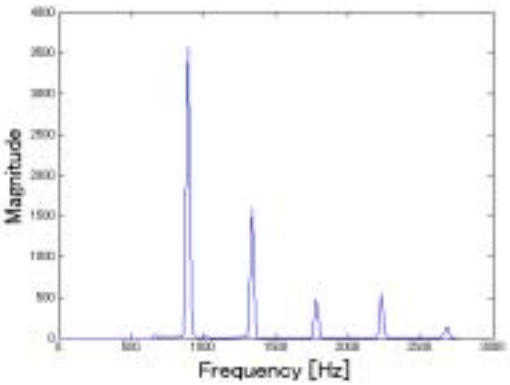

図３　ピアノの単音(A4)のスペクトル
(ミッシングファンダメンタル)

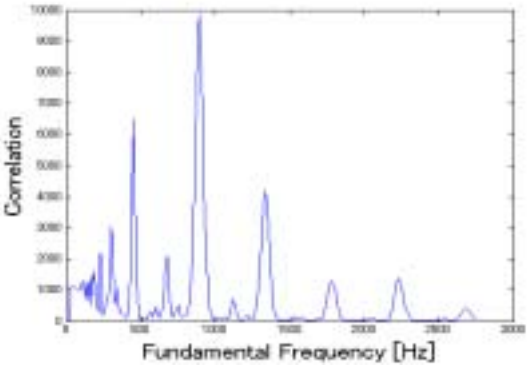

図４　ピアノの単音(A4)の相関
(ミッシングファンダメンタル)

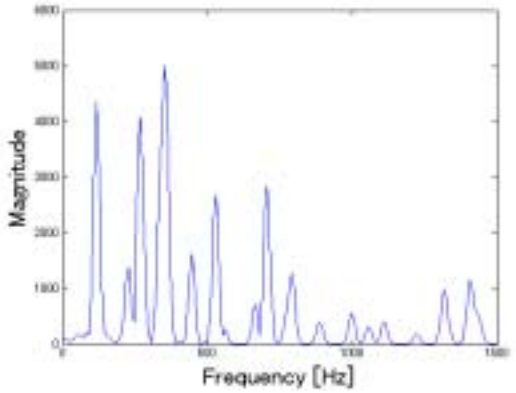

図 5　ピアノの混合音(音源数 3)のスペクトル
(音源名：A2、C4、F4)

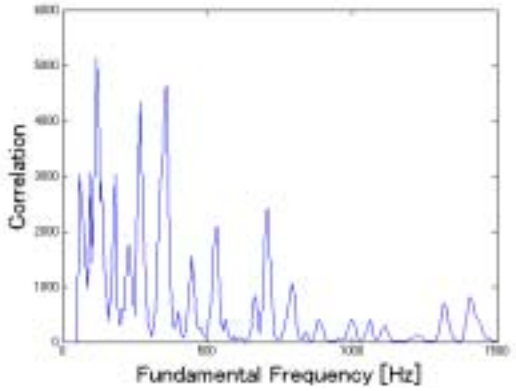

図 6　ピアノの混合音(音源数 3)の相関
(音源名 A2、C4、F4)

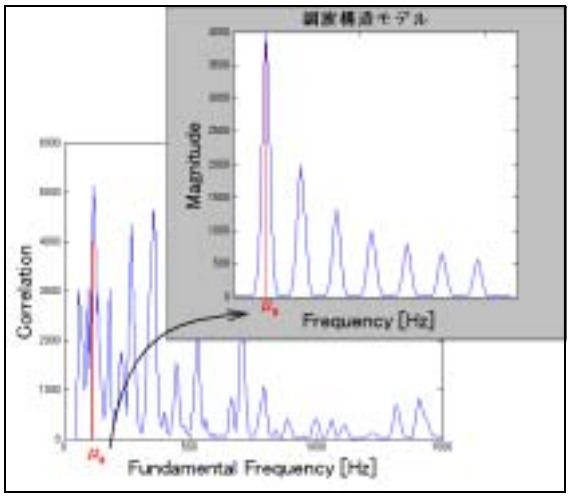

図 7　モデルの振幅値の決定

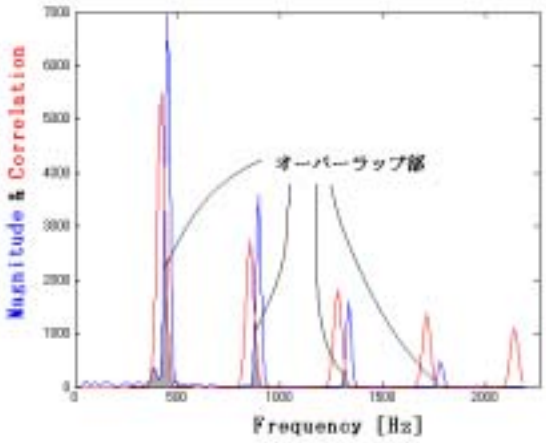

図 8　類似度の概念図

4. 相関を用いた類似度の定式化

4.1. 相関による手法の検討と発展

調波モデルと観測スペクトルとの相関を算出し、そのピーク抽出を行うことで基本周波数の推定を行えることを示した。しかし、図 6 に見られるとおり、ある周波数のハーフピッチや低い周波数に極大値が多く、かつそれらの値が大きい。本節では 1 フレームでの基本周波数推定のみに特化し、音源数が与えられているときの性能の向上するための検討をする。

4.2. 評価関数の類似度への拡張

調波構造モデルと観測スペクトルの相関を計算する際、モデルは調波構造を有しているが、最大値は 1 であり､振幅の絶対値としての情報はない。振幅値を定数倍しても相関値も定数倍されるのみである。そこで各基本周波数相応の振幅値を、基本周波数 μ_0 に相当する相関値 C(μ_0)を振幅値に採用する (図 7)。また、(4)式のような類似度関数を新たに定義する。

$$S(\mu)=\sum_{i=0}^{M}\begin{cases}\varphi(\omega_i,\mu), & (\varphi(\omega_i,\mu)<f(\omega_i))\\ f(\omega_i), & (\varphi(\omega_i,\mu)>f(\omega_i))\end{cases} \quad (4)$$

式(4)は調波構造モデルと観測スペクトルとの類似度を示す評価関数であり、周波数領域においてモデルとスペクトルが重なった部分の面積を求めていることに相当する(図 8)。

処理の流れの中で、毎フレーム類似度関数から音源数分のローカルピークを抽出する。また、評価式としては(4)式を用いるが、高調波になるほど振幅の値が無視できるため、モデルの倍音の数を有効であると考えられる倍音までで打ち切る。そして、各調波の面積を同等に評価を行うために、各調波に重みをつけて評価する。今挙げたパラメータは経験的に決めた定数である。図 9 に最終的な処理の流れを示す。

5. 適用例

5.1. 実験諸元

ピアノの単音を混合させた音源を対象にし類似度関数を適用した例を以下に示す。

周波数分析は短時間フーリエ変換を行い、サンプリング周波数は 24[kHz]、フレーム長は 150[msec],フレームシフトは 10[msec]で窓関数

にはガウス窓に 2 階のカーディナル B－スプライン関数を畳み込んだ時間窓を採用した[9]。また、FFT ポイントは 4096 とした。この窓は周波数領域で μ_0 を頂点とした山となり、その隣接調波位置以上/以下の値はゼロとなり他調波への漏れはない。

音源には RWC 研究用音楽データベース[10]の楽器音から楽器 No.1、バリエーション番号１から単音を抜き出して、適当に混合させた二つの音源を対象にする。

5.2. 実験データ

混合させた混合音データは以下のとおりである。

○音源１
音源数：３（A2、C4、F4）
演奏時間：2.6[sec]
○音源 2
音源数：5（B2、E3、C4、A4、D#5）
演奏時間：2.6[sec]

実験結果として、全部フレームに対し抽出された基本周波数を最も近い音名に量子化し、正解率を計算した(表 1、2)。また、音源 1、2 ともに各フレームごとに推定した周波数を図 10、11 に示す。ただし、各音源の立ち上がり、立下りなどがない安定した 200 フレームである。各単音の基本周波数を表 3 に示す。

5.3. 実験結果

表 1　音源 1 の正解率

音名	A2	C4	F4
正解率[%]	74.9	11.7	56.5

表 2　音源 2 の正解率

音名	B2	E3	C4	A4	D#5
正解率[%]	95.4	45.6	93.9	78.7	6.4

正解率を見ると、音源 1 では A2、音源 2 では B2、C4、A4 が安定して推定できている。しかし、他の音に関しては悪い結果となった。特に、音源 2 の D#5 がほとんど推定できていない。その様子を図 10、11 で確認してみると、フレーム単位で系列が上下しているが、推定できている音源に関してはストリームとして視覚的に捉えることができる。また、2 つの音源とも低周波数(60Hz 近辺)に推定誤りがある。

推定精度が悪かった理由として周波数分解能が原因の一つに考えられる。分解能を向上ことによりより正確に観測スペクトルを得られるからである。また、低周波に推定誤りがあったのは、観測スペクトルにはない半ピッチが相関値では出るので、その影響と考えられる。

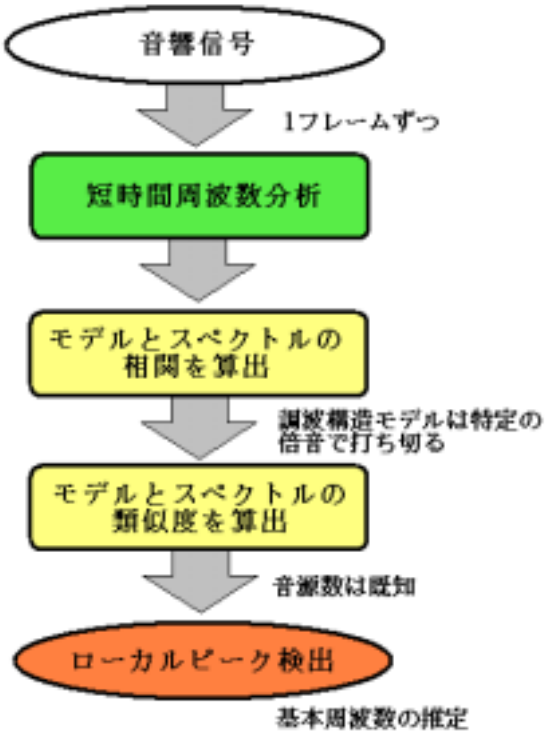

図 9　基本周波数推定までの処理の流れ

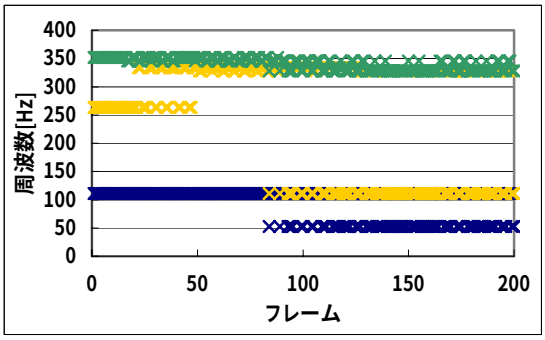

図10　基本周波数の推定結果(音源1)

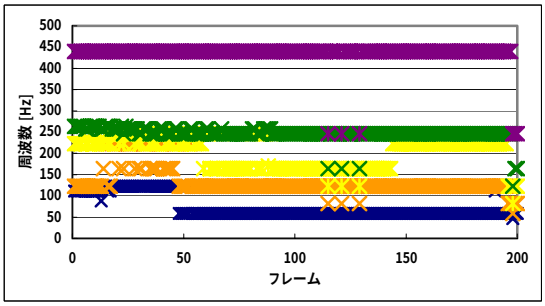

図11　基本周波数の推定結果(音源2)

表3　実験で用いた単音の基本周波数

音名	A2	C4	F4	
周波数[Hz]	110.0	261.6	349.2	
音名	B2	E3	A4	D#5
周波数[Hz]	123.5	164.8	440.0	622.3

6. おわりに

本報告では、音源分離再合成システムの構築を目的とし、混合音の音響信号から基本周波数を推定する問題を検討した。まず、周波数領域において調波構造のモデル化を行った。また、そのモデルと観測スペクトルの相関、類似度によって評価関数を計算し、ローカルピーク抽出によって基本周波数の推定をした。実験を行った 2 つの音源に関して、一部は安定して推定できたが、ほとんど推定できない音源があり、原因の追究とアルゴリズムの拡張による推定精度の向上を目指す。

今回は実験データも少なく、大量にデータを取り提案手法をより洗練させる必要がある。また、評価の面でも定量的に評価を行う必要がある。

今後の検討として、十分に評価実験を行うこともあるが、提案手法を洗練させる理由で周波数分解能の向上が必要である。これにはマルチレートフィルタバンク、もしくはウェーブレット変換などう分析手法の改善を考えている。また、展望として分離再合成という枠組みで、本手法を利用したいと考えている。

参考文献

[1] 三輪多恵子, 田所嘉昭, 斎藤努: “くし形フィルタを利用した採譜のための異楽器音中のピッチ推定”電子情報通信学会論文誌, Vol. J81-D-II, No. 9, pp. 1965–1974 (1998).

[2] 柏野邦夫，中臺一博，木下智義，田中英彦, “音楽情景分析の処理モデル OPTIMA における単音の認識,” 電子情報通信学会論文誌, vol.J-79-D-II, no.11, pp.1751-1761 (1996).

[3] 柏野邦夫, 木下智義, 中臺一博, 田中英彦, “音楽情景分析の処理モデル OPTIMA における和音の認識,” 電子情報通信学会論文誌, vol.J-79-D-II, no.11, pp.1762-1770 (1996).

[4] 後藤真孝, “音楽音響信号を対象としたメロディーとベースの音高推定, ”電子情報通信学会論文誌, D-Ⅱ, Vol. J84-D-Ⅱ, No.1,pp. 12–22 (2001).

[5] M. Goto: “A Predominent-F0 Estimation Method for Realworld Musical Audio Signals: MAP Estimation for Incorporating Prior Knowledge about F0s and Tone Models,” Proceedings of CRAC-2001, (2001).

[6] M. Goto: “A Predominent-F0 Estimation Method for CD Recordings: MAP Estimation Using EM lgorithm for Adaptive Tone Models,”Proceedings of the 2001 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2001, pp. V-3365–3368 (2001).

[7] 亀岡弘和, 西本卓也, 嵯峨山茂樹, “ガウス基底音響ストリームモデルを用いた時空間クラスタリングによる多重スペクトル分離,” 日本音響学会 2005 年春季研究発表会講演論文集, 3-7-19, pp.601-602 (2005).

[8] 亀岡弘和, 嵯峨山茂樹, “EM アルゴリズムと多重音解析への応用,” (チュートリアル講演), 日本音響学会音楽音響研究会資料 (2005).

[9] 河原英紀，片寄晴弘，Roy D. Patterson, Alain de Cheveigne，“瞬時周波数を用いた基本周波数の高精度の抽出について，”日本音響学会聴覚研究会資料，H-98-116, pp.31-38 (1998).

[10] 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: 音楽ジャンルデータベースと楽器音データベース,” 情報処理学会 音楽情報科学研究会 研究報告 2002-MUS-45-4, Vol.2002, No.40, pp.19-26 (2002).