逐次的な調波時間構造化クラスタリングによる多重音ピッチ推定*

◎江頭幸路, 宮本賢一, 小野順貴, 嵯峨山茂樹 (東大・情報理工)

1 はじめに

モノラル音楽音響信号中の多重音の各ピッチ（基本周波数）を推定する技術は、自動採譜や音楽加工・音楽検索などに幅広く有効な基礎技術である。当研究室ではそのための手法として調波時間構造化クラスタリング（Harmonic-Temporal-structured Clustering: HTC）[1] を開発し、その有効性を発表してきた。HTC は入力音響信号全体のスペクトログラムに対して、多数の単音エネルギー分布モデル（Fig. 1(a). 以下、単音モデル）をうまく当てはめることにより、信号中の各音のピッチや発音時刻などを一括で推定する。

一方で、人間の聴覚は知覚した音に対して逐次そのピッチを認識している。HTC でも逐次推定が可能になれば、分析したい音楽の終了を待たずにピッチ推定の結果を得られるようになる。以上の動機から、本研究では HTC を逐次的に行うことにより、音響信号を入力し次第、随時ピッチ推定する手法を示す。

2 逐次推定処理の定式化

2.1 問題設定

人間の聴覚は、人が意識して音楽的な構成を考慮したりパターンに基づく予測を行う場合を除いて、短時間の音の知覚結果を元にピッチを認識していると考えれらる。しかし Bregman の分凝要件 [2] によれば、まとまった音響エネルギー分布を 1 つの音として知覚するには音響エネルギーの時間変化が滑らかでなければならない。したがってその滑らかさを判断できる程度の時間が必要である。本研究の問題は、エネルギー変化の滑らかさが判断できる程度に短時間のスペクトログラムに対し、単音モデルを効果的にフィッティングしていく手法の開発である。

以上の問題を解決するために、HTC は次の 2 点に対応する必要がある。1 つは鳴っている音の数が各時刻で変化するためモデルの位置だけでなくフィッティングに必要な単音モデル数も変動すること、もう 1 つは発音・消音時刻がモデルフィッティングする時間区間内に含まれない状況が多く発生することである。

2.2 アルゴリズムの実現方針

どの時刻においても音響エネルギーの時間変化の滑らかさを等しく測れるようにするには、入力スペクトログラム上を少しずつ移動する短時間の分析窓（Fig. 2）を用意し、その窓の中でモデルを徐々にフィッティングしていけばよいと考えられる。2.1 節で示した単音モデル数変動の問題は、過剰なモデル数を自動調節する手法 [3] を利用できる。つまり、移動して分析窓に入ってきたばかりのスペクトル上に余分に単音モデルを置き、単音モデルを変形・移動させつつ実際に鳴っている音がある位置のモデルのみを推定過程で残す。モデルフィッティングは [3] で述べられているように、コスト関数の最小化によって達成できる。

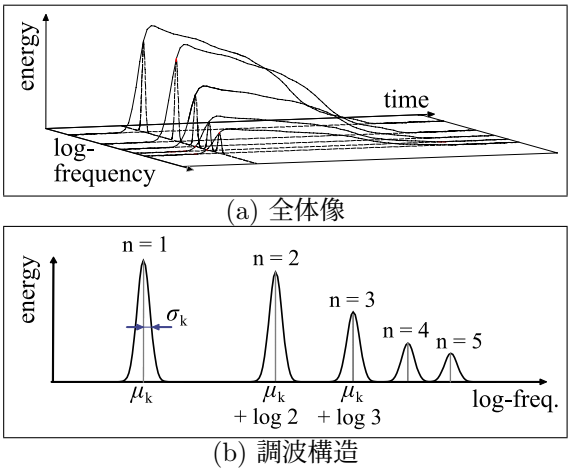


Fig. 1 単音の音響エネルギー分布モデル

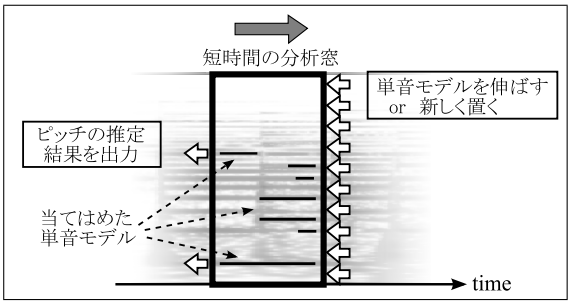


Fig. 2 スペクトログラム上を移動する分析窓

2.3 単音エネルギー分布のモデリング

単音モデルは、周波数方向については [1] と同様、Fig. 1(b) のように倍音間隔拘束付きの混合ガウスモデルを用いる。時間方向については、2.1 節で示した 2 つ目の問題があり、分析窓より長い音長のモデルを容易に扱えるようにするためにノンパラメトリックにモデリングする。そうすると時間方向のエネルギー包絡が滑らかとは限らなくなるので、滑らかさは目的関数でコストとして扱うことにする。したがって、k 番目の単音モデルが対数周波数-時間平面上の位置 (x,t) で持つエネルギーの大きさ $q_k[x,t]$ は

$$q_k[x,t] \stackrel{\text{def}}{=} w_k[t] \sum_n \frac{v_{k,n}}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x-\mu_k-\log n)^2}{2\sigma_k^2}\right) \tag{1}$$

と表せる。モデルパラメータはそれぞれ、$w_k[t]$：時刻 t でのモデルのエネルギー、v_{kn}：n 番目の倍音成分の相対的大きさ ($\forall k\colon \sum_n v_{kn}=1$)、$\sigma_k$ ：各倍音成分のエネルギーの周波数方向の分散、μ_k：その単音のピッチ、である。

---

* "Sequential Harmonic-Temporal-structured Clustering for Multipitch Estimation in Polyphonic Music Signals" by Koji EGASHIRA, Kenichi MIYAMOTO, Nobutaka ONO, and Shigeki SAGAYAMA, Graduate School of Information Science and Technology, The University of Tokyo.

2.4　最小化する目的関数の設計

入力スペクトログラム $W(x,t)$ と単音モデルの混合エネルギー分布 $\sum_k q_k[x,t]$ の類似度はそれらの I ダイバージェンスにより測ることができる。余分なモデルの削除のために、[3] 同様のペナルティ項を目的関数に加える。また、単音の時間方向のエネルギー包絡が滑らかになるように、隣り合う時間フレームでのモデルエネルギーの差の二乗和をコストとして設定した。以上から、最小化する目的関数はこれら 3 つのコストの総和として以下のように設定できる。

$$
\begin{aligned}
J \stackrel{\mathrm{def}}{=} & \sum_{x,t}\{W[x,t]\log\frac{W[x,t]}{\sum_k q_k[x,t]} \\
& -(W[x,t]-\sum_k q_k[x,t])\} + \chi \sum_{\substack{\forall\, k,t \text{ s.t.} \\ w_k[t]>0}} \log w_k[t] \\
& +\phi\sum_{k,t}(w_k[t+1]-w_k[t])^2
\end{aligned}
\tag{2}
$$

第 1 項が I ダイバージェンス、第 2 項がモデル削除のための項、第 3 項がエネルギー包絡の滑らかさのための項であり、χ,ϕ は I ダイバージェンスに対するその他の項の重みである。

第 2 項はなるべく少ないモデルで入力スペクトログラムを表現しようとする働きを持つ [3] ため、音響エネルギーが少数のモデルに集中し、多数のモデルのエネルギーは無くなる。エネルギーの無くなったモデルが再びエネルギーを持つことは無く、またそのようなモデルは音の存在を表さないため無意味なので、削除できる。

2.5　補助関数を用いた目的関数の最小化

(2) 式はそのままでは最小化が難しいため、補助関数を利用して最小化する。[1] や [4] で用いられている補助関数を (2) 式に適用することで、(2) 式を上から押さえる補助関数を用意できる。この補助関数を偏微分することにより、補助関数の値を小さくする各モデルパラメータと補助変数の更新式を導出でき、各モデルパラメータと補助変数を繰り返し更新することにより (2) 式は局所最小に収束する。

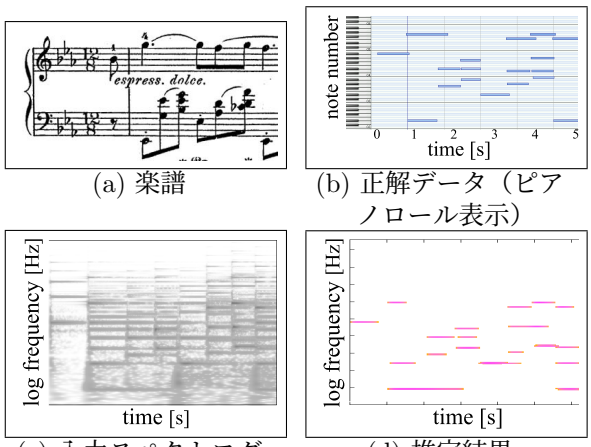
2.6　逐次推定アルゴリズム

本手法のアルゴリズムは以下のようにまとめられる。

1. 入力音響信号のスペクトログラムを計算する。
2. 分析窓を入力スペクトログラムの先頭に置く。
3. 分析窓に入ってきた部分のスペクトログラム上に新しく単音モデルを配置する。
4. (2) 式を小さくするように、各モデルパラメータと補助変数を更新する。エネルギーが無くなったモデルは削除する。
5. 分析窓を数フレーム移動する。分析窓から出た時刻のモデルパラメータは更新を終了する。
6. 分析窓が入力スペクトログラムの最後まで到達したら推定処理は完了。そうでなければ、分析窓内にて鳴り終わっていない音のモデルを分析窓が移動した分だけ伸ばし、3 に戻って繰り返す。

3　音楽音響信号による実験評価

提案手法の性能を確かめるため、プログラムに実装し、RWC 音楽データベース収録の実演奏データに対して適用し推定結果の正解率を算出した。用いたデータはクラシック 1 曲（RWC-MDB-C-2001 No.30）、ジャズ 9 曲（RWC-MDB-J-2001 No.{1-3,6-10,13}）の計 10 曲の冒頭 30 秒ずつで、推定結果のうち音長が 200ms 以下のものは削除した。正解判定は時間フレームごとではなく単音ごとに行った。



(a) 楽譜　(b) 正解データ（ピアノロール表示）
(c) 入力スペクトログラム　(d) 推定結果

Fig. 3　ショパン作 ノクターン Op.9-2 の冒頭 5 秒間への提案手法の適用結果

10 曲の平均正解率は、Precision: 43.5%、Recall: 63.5%であった。例として、ショパン作曲のノクターン Op.9-2（RWC-MDB-C-2001 No.30）に対する実験結果を Fig. 3 に示す。多くの単音においてピッチが正しく推定できているが、推定誤りの原因として、音の存在を検出できなかったり倍音成分の周波数にピッチが推定されたものが見受けられる。

4　おわりに

本研究では、人間の聴覚が音のピッチを逐次認識していることをヒントに、スペクトログラム上を移動する分析窓内で調波時間構造化クラスタリングを行い、多重音のピッチ推定を逐次的に行う手法を示した。今後の発展として、特に音の鳴り始めを意識した、単音エネルギー分布モデルの改良を検討している。

本研究の一部は科学研究費補助金・基盤研究A (課題番号 00303321) の補助を受けて行なわれた。

参考文献

[1] H. Kameoka *et al.*, “A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering,” *IEEE Trans. Audio Speech Lang. Proc.*, 15 (3), 982-994, 2007.
[2] A. Bregman, Auditory Scene Analysis, MIT Press, 1990.
[3] 江頭他、“クラスタ数を自動調整する調波時間構造化クラスタリングによる多重音ピッチ推定,” 日本音響学会春季研究発表会講演集, 899-900, 2008.
[4] N. Ono *et al.*, “Separation of a Monaural Audio Signal into Harmonic/Percussive Components by Complementary Diffusion on Spectrogram,” *Proc. of EUSIPCO*, 2008