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*

◎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

本研究で扱うモノラル音響信号からの多重音ピッチ推定は、自動採譜や音楽加工、音楽検索などへの応用が期待されて長年研究されているテーマであるが、未知である情報が多いことなどを理由にいまだに難しい問題である。この問題に対して、多くの従来手法は、まず各時刻ごとに音の倍音構造を利用してピッチ推定した後に、音の時間的つながりを考慮して最終推定結果を得る段階的手法を採ってきた。

これに対し我々は、音の周波数構造と時間構造を推定過程上分けずに同時に扱う調波時間構造化クラスタリング（Harmonic-Temporal Clustering; HTC）を提案してきた [1]。HTC は Bregman の分凝要件と呼ばれる心理物理実験からの知見 [2] に基づいて、単音の持つエネルギーの倍音構造と時間変化の倍音成分間類似性や滑らかさといった構造の両方を備えたパラメトリックなエネルギー分布モデル（以後単音モデルと呼ぶ）を用意し、そのモデルに合うように観測信号の音響エネルギーを単音ごとにクラスタリングすることにより、従来手法より高い推定精度を得ることができる。

本研究では、推定結果の初期値依存性の緩和などによりさらなる性能向上を目的として、クラスタ数の自動調整機能を組み込んだ HTC について論じる。

2　調波時間構造化クラスタリング

まず本研究が利用する既存手法 [1] の概要を述べる。観測音響信号を短時間周波数分析して得られるスペクトログラムを $W(x,t)$（x：対数周波数、t：時刻）で表す。多数の音が様々なピッチ・時刻・音長・音量で鳴っている観測信号のスペクトログラムは各音のスペクトログラムの重ね合わせで表せるという仮定のもとで、$W(x,t)$ を多数の単音モデルでフィッティングし一音ごとにクラスタリングして、それぞれのモデルのパラメータから各単音のピッチや発音時刻などを得る。

スペクトログラムで表される単音モデルを $q_k(x,t)$ $(k=1,\ldots,K, K$：モデルの総数) で表すと、$W(x,t)$ と $q_k(x,t)$ の全モデルの和との間の分布間距離は Kullback-Leibler (KL) ダイバージェンスを用いて

$$J_o \stackrel{\text{def}}{=} \iint W(x,t)\log\frac{W(x,t)}{\sum_k q_k(x,t)}dxdt$$

$$\text{ただし}\iint W(x,t)dxdt = \iint\left(\sum_k q_k(x,t)\right)dxdt \tag{1}$$

と表現できる。$q_k(x,t)$ は周波数方向の倍音構造と時間方向の連続的包絡構造を成すように拘束された 2 次元混合ガウス分布

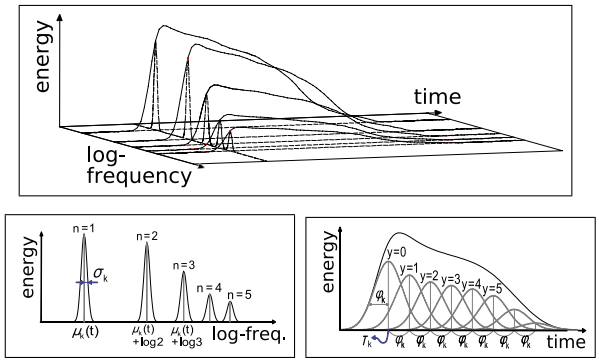


Fig. 1　単音モデル $q_k(x,t)$：(上) 概形、(左下) 倍音構造、(右下) エネルギーの時間変化

$$q_k(x,t) \stackrel{\text{def}}{=} \sum_n\sum_y \frac{w_k v_{kn} u_{ky}}{2\pi\sigma_k\phi_k} e^{\left(-\frac{(x-\mu_k-\log n)^2}{2\sigma_k^2}-\frac{(t-\tau_k-y\phi_k)^2}{2\phi_k^2}\right)},$$

$$\sum_n v_{kn} = \sum_y u_{ky} = 1 \tag{2}$$

で表現できる (Fig. 1)。パラメータ w_k, μ_k, τ_k はそれぞれ音量、ピッチ、発音時刻を意味し、$\sigma_k, \phi_k, v_{kn}, u_{ky}$ は、各ガウス分布の周波数方向の標準偏差、時間方向の標準偏差、周波数方向の各ガウス分布間の大きさの相対比率、時間方向の相対比率である。

3　モデル数調整を行う HTC

3.1　モデル数推定の必要性と初期値依存性

ここで HTC をクラスタリングの観点から見直す。クラスタリングには一般に 2 つの大きな問題が知られており、1 つは最適なクラスタ数の決め方である。HTC の場合は良い推定結果を得るためには、実際に鳴っている単音数と同数程度のクラスタが存在することが望ましく、クラスタが多すぎると推定結果に余計な音が加わるようになり、逆に少なすぎると本当は鳴っているはずの音が推定結果から漏れる傾向がある。しかし観測信号中で鳴っている単音数は不明なので、クラスタリングと合わせて単音数も推定する必要がある。広く利用される対処法として AIC や BIC、MDL などのモデル数選択規準が用いられる。

もう 1 つはクラスタリング結果に対するモデルパラメータの初期値依存性であり、この問題についてはランダム初期化による複数回試行や入力サンプルのヒストグラムに基づく初期化などの対策がしばしば用いられる。

これら 2 問題を個別に解決する手法は既に挙げたように存在するが、本研究では 2 問題を同時に解決

---

* "Unsupervised Adjustment of the Number of Clusters in Harmonic-Temporal Clustering for Multipitch Estimation," by Koji EGASHIRA, Kenichi MIYAMOTO, Nobutaka ONO, and Shigeki SAGAYAMA, Graduate School of Information Science and Technology, The University of Tokyo.

する手法を検討する。

3.2 スパース化に基づいたモデル数調整

単音モデル数を調整するとき、数を増やす操作は新しく追加したモデルをどこに配置すれば良いのか分かりにくいことなど、扱いづらい点が多い。そこでモデル数を減らして調整するために、単音モデルのスパース化の考え方を導入する。

推定結果の単音モデル初期配置への依存性に対処するために、HTC 初期化時に観測信号中に音のエネルギー分布が存在する可能性のある全ての位置に単音モデルを配置し全ての w_k に等しく値を与えておき、その上でなるべく少ないモデルで観測信号を表現するためにモデルの音量パラメータ w_k をスパース化する。w_k のスパース性を表すコスト関数に $\sum_k \log w_k$ を用い、重み χ をかけてペナルティ項として (1) に付加したもの：

$$J_m \stackrel{\text{def}}{=} J_o + \chi \sum_k \log w_k \tag{3}$$

を最小化する目的関数とする定式化ができる。

(3) の最小化を最適化問題として解くために [1] と同様に各 x, t における $W(x,t)$ を各単音モデルに分配する変数 $m_k(x,t)(\sum_k m_k(x,t) = 1)$ を導入し、$m_k(x,t)$ の更新とモデルパラメータの更新を交互に行う反復推定により (3) は単調減少し局所最適解に収束する。w_k の更新式は $L_k \stackrel{\text{def}}{=} \iint m_k(x,t)W(x,t)dxdt$ として

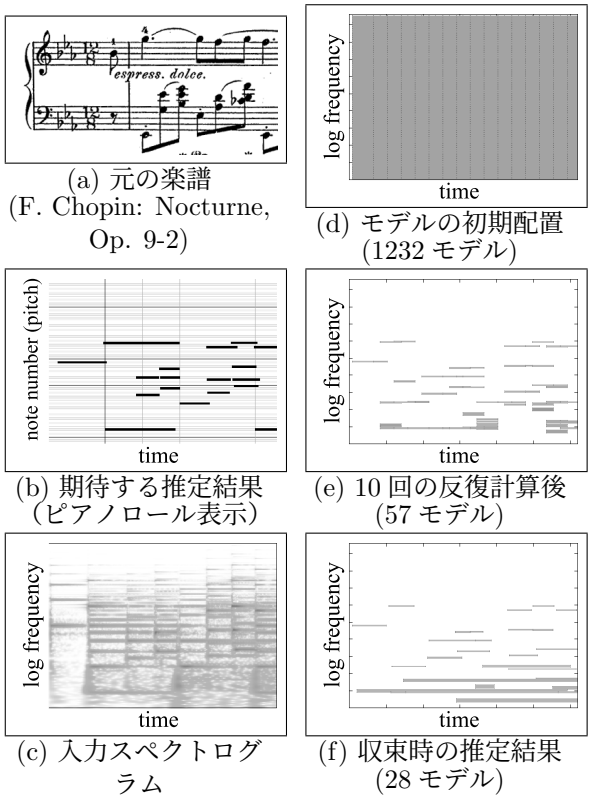
$$w_k = \begin{cases} \dfrac{(L_k - \chi)\iint W(x,t)dxdt}{\sum\limits_{\forall k \mid L_k > \chi} (L_k - \chi)} & \text{if} \quad L_k > \chi \\ 0 & \text{if} \quad L_k \le \chi \end{cases} \tag{4}$$

となり、この更新によって $w_k = 0$ となったモデルは $W(x,t)$ へのフィッティングに寄与しないので削除できる。またこの式 (4) から、χ の大きさは単音モデルが持つべき最小エネルギーに決めればよいと考えられる（他のパラメータや $m_k(x,t)$ の更新式は [1] 参照）。この定式化は [3] で示されている手法と似たものになっている。

4 実装システムの適用例

提案アルゴリズムを実装し、実際の楽曲演奏を録音した音響信号に適用した例を示す。対象は RWC 研究用音楽データベース収録のピアノ楽曲、ショパン作曲ノクターン No.2, Op. 9-2 の冒頭約 5 秒間であり、この区間に単音は 18 個含まれている。なおピアノの打鍵による非調波的エネルギーにモデルが多数フィットするのは本手法の趣旨から外れるため、[4] の手法を用いて非調波成分を抑制した信号を入力に用いた。単音モデルの初期配置は時間周波数領域全面を覆うように各音階に相当する周波数ごと、300 ms 間隔で行った。

Fig. 2 に示す実験結果から、本手法によって余分な単音モデルが削除されつつピッチ推定が行えることが分かる。しかし正解の演奏情報 (Fig. 2(b)) と比べると特に低音階部において実際には鳴っていない音の位置にモデルが残っている。その理由は少ないモデルで観測信号を表現しようとするためにモデルを低音部に配置しそのモデル 1 つの倍音成分で高音部において実際に鳴っている複数の単音を表現しようとするからだと考えられる。



(a) 元の楽譜 (F. Chopin: Nocturne, Op. 9-2)
(d) モデルの初期配置 (1232 モデル)
(b) 期待する推定結果（ピアノロール表示）
(e) 10 回の反復計算後 (57 モデル)
(c) 入力スペクトログラム
(f) 収束時の推定結果 (28 モデル)

Fig. 2 実演奏データに対する推定結果

5 おわりに

本研究では、HTC のクラスタリングにおける最適クラスタ数決定と初期値依存性の 2 問題に対処するために HTC にクラスタ数調整機能を組み込んだ手法を提案し、実演奏の録音データへの適用例を示した。今後の課題として、小音量や短音長の音に対してモデルが削除されやすい問題の克服などを検討している。

謝辞 本研究の一部は科学研究費補助金・基盤研究 B (課題番号 17300054) 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] M. Figueiredo *et al.*, “Unsupervised Learning of Finite Mixture Models,” *IEEE Trans. Pat. Anal. Mach. Intell.*, 24 (3), 381-396, 2002.
[4] 宮本他, “スペクトログラム 2 次元フィルタによる調波音・打楽器音の分離,” 音講論 (秋), 825-826, 2007.