確定的アニーリング EM アルゴリズムを用いた調波時間構造化クラスタリングによる音楽信号分析*

◎亀岡弘和, 西本卓也, 嵯峨山茂樹 (東大情報理工)

1 はじめに

本報告では、我々が以前提案した多重音解析法である調波時間構造化クラスタリング (Harmonic-Temporal-structured Clustering; HTC)[1] におけるパラメータ反復推定部において、EM アルゴリズムを適用した場合と、EM 版の焼きなまし法である確定的アニーリング (DA)EM アルゴリズム [3] を適用した場合の性能比較を行う。HTC は、時間周波数平面上に「静的」に分布するスペクトルを観測パターンと捉え、スペクトルの時間と周波数方向の大域構造を 2 次元モデリングし、一挙に音高、オンセット、音長を推定するものである。この考え方は、フレームごとのスペクトルを観測パターンと捉え、複雑な「動的」推定問題と考える従来のアプローチとは根本的に異なるものである。文献 [1] ではパラメータ推定を EM アルゴリズムとして定式化したが、大域最適解に収束させるにはうまく初期値を与える必要があり、局所収束性の問題が懸案課題の一つとして残されていた。そこで、EM アルゴリズムにおける局所収束性の問題に対し、効果的な対応策として最近開発された DAEM アルゴリズムを HTC に適用し、どの程度性能向上に貢献するかを評価した。

2 調波時間構造化クラスタリング (HTC)

ここでは、HTC の定式化の概略を述べる (定式化の詳細は文献 [2] 参照)。HTC は、周波数方向の調波構造と時間方向の連続的なエンベロープをなす k 番目の単一音のスペクトルパターンを 2 次元幾何分布モデル $q_k(x,t;\Theta)$(音響オブジェクトモデル k と呼ぶ) の係数 w_k の重みつき和モデル (モデルパラメータは Θ) で観測スペクトル $W(x,t)$ を、最も良く近似する方法論である。ただし、x、t は対数周波数と時間である。音響オブジェクトモデルは、

$$q_k(x,t;\Theta) = w_k \sum_{n=1}^{N} \frac{v_{kn} u_k(t)}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k-\log n)^2}{2{\sigma_k}^2}} \quad (1)$$

$$u_k(t) = \sum_{y=0}^{Y-1} \frac{u_{ky}}{\sqrt{2\pi}\phi_k} \exp\left(-\frac{(t-\tau_k-y\phi_k)^2}{2{\phi_k}^2}\right) \quad (2)$$

で与えられる。各パラメータの詳細を図 1〜3 に示す。このモデル関数は以下のような和の形

$$q_k(x,t;\Theta) = \sum_{n=1}^{N}\sum_{y=0}^{Y-1} S_{kny}(x,t;\Theta) \quad (3)$$

に書き直すことができる。EM アルゴリズムでモデル推定を行う場合、M ステップの Θ の更新式は

$$\widehat{\Theta} = \operatorname{argmax} \sum_{\forall k}\left(\left(\sum_{\forall n}\sum_{\forall y}\iint_D m(k,n,y;x,t) W(x,t)\log S_{kny}(x,t;\Theta)\,dx\,dt\right) - d_v\sum_{\forall n}\bar{v}_n\log\frac{\bar{v}_n}{v_{kn}} - d_u\sum_{\forall n}\sum_{\forall y}\bar{u}_y\log\frac{\bar{u}_y}{u_{kny}}\right) \quad (4)$$

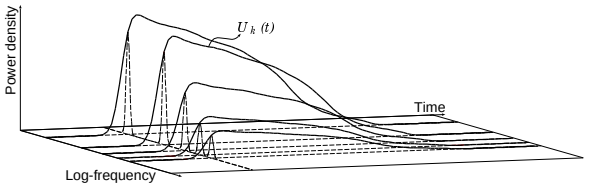

Fig. 1 音響オブジェクトモデル k

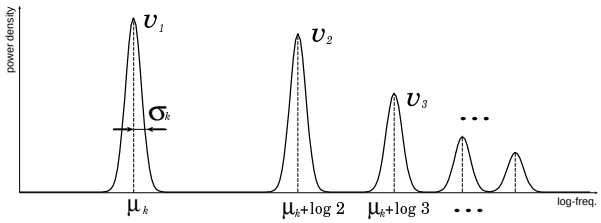

Fig. 2 時刻 t における $q_k(x,t;\Theta)$ の切口 (調波構造)

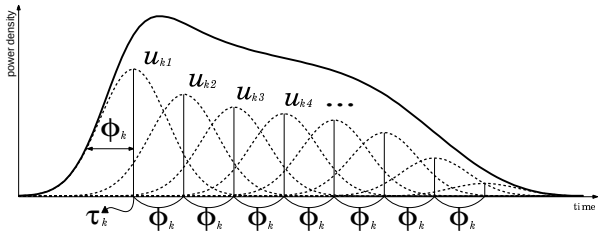

Fig. 3 パワーエンベロープ関数モデル $u_k(t)$

から導出すれば良く (Lagrange 未定乗数項は省略してある。$\bar{v}_n$、d_v、$\bar{u}_y$、d_u は v_{kn}、u_{ky} に関する事前分布項に関係する実験的に定める定数である。)、エネルギー分配関数 $m_{k,n,y}(x,t)$ は E ステップにおいて

$$m_{k,n,y}(x,t) = \frac{S_{k,n,y}(x,t;\Theta)}{\sum_{\forall k}\sum_{\forall n}\sum_{\forall y} S_{k,n,y}(x,t;\Theta)} \quad (5)$$

に更新すれば、反復計算の末、局所最適解に収束することが保証される。

3 DAEM アルゴリズム [3]

EM アルゴリズムは解の停留点への収束が保証されているものの、初期値の選び方によっては必ずしも大域最適解に収束しない。上田らが開発した DAEM アルゴリズム [3] は、EM アルゴリズムに焼きなまし法を応用して局所解に陥らないように工夫した手法である。上田らは、EM アルゴリズムの局所収束が悪い初期値の影響を陽に受けるのは、反復計算の早期において内部状態事後確率 (HTC においてはエネルギー分配関数に対応) の信頼性が低いためであり、早期の反復計算時には、$W(x,t)$ は $\{k,n,y\}$ 番目のクラスタ (ここで言うクラスタとは $m_{k,n,y}(x,t)W(x,t)$ をさす) にその時のパラメータに依らずほぼ平等に分配するべきであると考えた。そしてこれは、多峰的な目的関数をなまけさせる効果になっている。式 (5)

---

* Music Signal Analysis based on Harmonic-Temporal-structured Clustering using Deterministic Annealing EM Algorithm. by KAMEOKA, Hirokazu, NISHIMOTO, Takuya, SAGAYAMA, SHigeki (University of Tokyo)

の代わりに、β パラメータを用いて新たにエネルギー分配関数を

$$m_{k,n,y}(x,t) = \frac{S_{k,n,y}(x,t;\boldsymbol{\Theta})^{\beta}}{\sum_{\forall k}\sum_{\forall n}\sum_{\forall y} S_{k,n,y}(x,t;\boldsymbol{\Theta})^{\beta}} \quad (6)$$

として定義し、通常の EM 反復計算に β のループも加えた二重の反復計算でパラメータを更新していくのが DAEM アルゴリズムの概要である。β が 0 に近いほど目的関数は単峰的になり、これを徐々に上昇させていき、最終的に 1 にすることで元の目的関数を評価することになる。DAEM アルゴリズムを用いた HTC は以下にようにして実装される。

---

1. Set $\beta \leftarrow \beta_{min} (0 < \beta_{min} < 1)$
2. Set $\boldsymbol{\Theta}^{(0)}$, $i \leftarrow 0$
3. Iterate EM-steps until convergence:
   - E-step: $\boldsymbol{\Theta}^{(i)}$ を式 (6) に代入
   - M-step: $\boldsymbol{\Theta}^{(i+1)} \leftarrow$ 式 (4)
   - Set $i \leftarrow i+1$.
4. Increase β.
5. If $\beta < 1$, repeat from step 3; otherwise stop.

---

4 評価実験

HTC のパラメータ推定部で EM アルゴリズムと DAEM アルゴリズムを行った場合の性能比較を行うため、RWC 研究用音楽データベースの実音楽音響信号 (実験データを表 1 に示す) を対象に評価実験を行った。

$W(x,t)$ は、ガボールウェーブレット変換 (サンプリング周波数 16kHz、時間分解能 16ms、最低周波数 60Hz、周波数分解能 12cent) により解析した。解析する時間周波数平面の時間区間は連続する 80 フレーム (1.28s) とした。HTM の対数基本周波数およびオンセット時刻 $(\mu_k, \tau_k | k = 1, \cdots, K)$ のパラメータ初期値は、$W(x,t)$ の極大点のうちエネルギーの大きいものから 20 点抽出して、それらの時間周波数平面上での座標とした。また、音響イベントの推定総数は w_k が閾値より大きい HTM の個数とした。実験条件をまとめて表 2 に記す。β の初期値は 0.3 とし、1 ループごとに 0.1 ずつ増加させた。w_k の推定値が閾値以下の場合、k 番目の音は無音と判定することで発音数の推定を行うことにした。

推定した音高、オンセット、音長を音階およびフレーム単位に量子化して、推定音高のフレーム系列を作成し、推定音高系列と参照音高系列から音高正解率を DP(Dynamic Programming) に基づいた自動計算法により計算した。実装手順は紙面の都合上省略する。この自動計算では、置換誤りを脱落と挿入の二重の誤りと判断するため、場合によっては正解率が負となることがある。

閾値による音高候補のトランケーションでは、閾値の大きさに応じた挿入誤りと脱落誤りの数との間にはトレードオフがあるが、さまざまな閾値を試した中で最も良い音高正解率が両者の性能を反映した基準になっていると我々は考えた (すなわち、発音数の推定法が今後検討していくべき課題であることを暗に示している)。各データにおいて、試した閾値の中で最も良かった正解率結果を表 3 に示す。

Table 1 評価実験に使用した実音楽信号データ

	Title	Catalog number
d1	Crescent Serenade	RWC-MDB-J-2001 No. 9
d2	Lounge Away	RWC-MDB-J-2001 No. 8
d3	For Two	RWC-MDB-J-2001 No. 2
d4	Jive	RWC-MDB-J-2001 No. 6
d5	Three Gimnopedies no. 1	RWC-MDB-C-2001 No. 35
d6	Nocturne no.2, op.9-2	RWC-MDB-C-2001 No. 30

Table 2 実験条件

Sampling rate	16 kHz
frame shift	16 ms
mother wavelet	Gabor function
frequency resolution	12.0 cent
frequency range	60–3000 Hz
initial # of HTMs	20
# of partials: N	6
# of kernels in $U_k(t)$: Y	10
$\bar{\nu}_n$	$0.6547 \times n^{-2}$
$\bar{u}_y$	$0.2096 \times e^{-0.2y}$
d_ν, d_u	0.04
range of analyzing segment	80 frames (1.28 s)
# of analyzing segments	21 (total time: 24 s)

Table 3 実験結果 (%)

	EM	DAEM
d1	88.7	88.8
d2	63.0	83.8
d3	60.6	81.7
d4	62.0	84.4
d5	64.1	72.1
d6	55.2	75.9

実験結果より、DAEM アルゴリズムを EM アルゴリズムの代わりに用いた場合に、すべてのデータに対して性能向上が見られ、HTC において DAEM アルゴリズムが効果的であることが確認できた。また、この結果は、正しい解と HTC における目的関数における大域最適解がうまく対応していることを裏付けているとも解釈できる。

5 おわりに

我々が開発した多重音解析法 HTC におけるパラメータ推定を DAEM アルゴリズムと EM アルゴリズムで行う場合での性能比較を行い、DAEM アルゴリズムが HTC に効果的であることが確認できた。

謝辞 本研究に関しいつも有益な議論をして頂いている、後藤 真孝 氏 (産総研) に感謝する。

参考文献

[1] 亀岡, 西本, 嵯峨山, “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音スペクトル分離,” 音講論 (春), 3-7-19, pp. 601–602, 2005.

[2] 亀岡, 西本, 嵯峨山, “調波時間構造化クラスタリング (HTC) による音楽音響特徴量の同時推定,” 情処研報, 2005-MUS-61-12, 2005.

[3] N. Ueda, R. Nakano, “Deterministic annealing EM algorithm,” *Nueral Networks*, 11(2), pp. 271–282, 1998.