調波時間構造化クラスタリングによるCASAへのアプローチ

亀岡　弘和[†]　　ルルー・ジョナトン[†]　　小野　順貴[†]　　嵯峨山茂樹[†]

† 東京大学大学院情報理工学系研究科
E-mail: †{kameoka,leroux,onono,sagayama}@hil.t.u-tokyo.ac.jp

あらまし　人間の聴覚機能を計算機で実現しようという試みが活発に進められており、その枠組を総称して計算論的聴覚情景分析 (Computational Auditory Scene Analysis: CASA) と呼ぶ。近年この研究分野における興味の対象は、Bregman が指摘した分凝要件 [1] に基づく混合音分離法の実現にある。CASA における多くの従来手法の間で共通するのは、各時刻で独立に調波成分を見つけ出すための処理 (周波数方向の群化) と、抽出された調波成分特徴量の時系列を時間方向にスムージングする処理 (時間方向の群化) を多段処理的に行っている点である。しかしながら、より良い群化プロセスの実践のためには、これらは本来協調し合いながら行われるべきであり、個々の音源の時間周波数全域に渡ったスペクトル構造を一挙に推定できる方法論が不可欠であると我々は考える。本稿では、この観点から導かれる調波時間構造化クラスタリングと呼ぶ CASA のための新しいアプローチを提案する。
キーワード　計算論的聴覚情景分析, Bregman の分凝要件, 定 Q フィルタバンク, 調波時間構造化クラスタリング

Harmonic Temporal Structured Clustering: A New Approach to CASA

Hirokazu KAMEOKA[†], Jonathan LE ROUX[†], Nobutaka ONO[†], and Shigeki SAGAYAMA[†]

† Graduate School of Information Science and Technology, The University of Tokyo
E-mail: †{kameoka,leroux,onono,sagayama}@hil.t.u-tokyo.ac.jp

Abstract　There have been many studies on computational implementation of human auditory function. Such a framework is referred to as computational auditory scene analysis (CASA). The recent interest in this area of research is focused on the development of source separation algorithms based on the grouping cues suggested by Bregman [1]. In most of the conventional methods for CASA, stream segregation is performed through two successive procedures: identification of the harmonic components and temporal smoothing of the extracted harmonic component features. These procedures, however, should be done essentially in a cooperative way and this belief has led us to formulate a unified estimation framework for the two dimensional structure of time-frequency components. We propose in this paper a new approach for CASA called Harmonic Temporal Structured Clustering.
Key words　CASA, Bregman's grouping cues, constant Q filter bank, Harmonic Temporal Structured Clustering

1. 序　　論

様々な音が混在する中で目的音の情報を計算機にうまく分離推定させることは、ロボット聴覚、音声認識や自動採譜などの様々な魅力的なアプリケーションを実現する上で工学的に重要な要素技術である一方で、数理的には大変難しい問題である。しかし、我々人間は、そのような環境の中でも容易に特定の音だけを選択的に聴取することができることから、聴覚は能動的に外界を把握するための優れた機能をもつと考えられている。このように音を通じて環境を把握することは聴覚情景分析 (Auditory Scene Analysis: ASA) と呼ばれ、Bregman の著書 [1] により広く知られることとなった。この著書の中で Bregman は、知識を利用しない聴覚の低次の分離能力に関して、

I.　音響信号はスペクトログラムに似た要素に「分解」される
II.　同じ音源に由来する要素は「群化」されて音脈を形成する
III.　群化のされやすさ (分凝要件) は、(a) 周期性 (調波構造)、(b) 調波成分の共通の立上り時刻、(c) 調波成分の共通の周波数および振幅変化、(d) 成分の連続性、(e) 時間周波数の近接性、(f) 共通の音源位置などに関係する

ことを心理実験を通して示している [1]。近年このような聴覚機能を計算機で実現しようという試みが活発に進められており、その枠組を総称して計算論的聴覚情景分析 (Computational Auditory Scene Analysis: CASA) と呼ぶ。

最近の CASA の研究における興味の対象は、Bregman が指摘した以上の分凝要件に基づく混合音分離法の実現にある。すなわち、分解と群化のプロセスを、分凝要件に関係する物理量

を用いてアルゴリズムとして実現し、音脈の認識に有用な特徴量 (ピッチ周波数など) を抽出したり、目的音に相当する音脈の再構成を行うことが主な目的である。

このような背景のもと、Cooke [2]、Brown ら [3]、Ellis [4]、Fishbach [5]、中谷ら [6] は分凝要件に基づく混合音分離の手法を提案している。これらの手法の多くは規則記述的や人工知能的なアプローチをとるため、トップダウン式にさまざまな制約条件を導入しやすい点では優れるが、細分化された閾値判定処理で構成される煩雑なアルゴリズムにならざるを得ないのが難点である。一方、西ら [7]、鵜木ら [8]、安部ら [9], [10]、Wu ら [11] は、分凝要件を数学的な制約として定義し、それに基づく制約つき最適化問題として定式化を行っている。これらに代表される CASA の多くの従来手法の間で共通するのは、時間周波数分解による分解プロセスを経て、群化プロセスが、瞬時特徴量の尤度を計算するステップと、その尤度に基づいて瞬時特徴量の時間軌跡を Kalman フィルタなどのダイナミカルシステム、隠れマルコフモデル、マルチエージェントシステムなどを用いて逐次的に追跡するステップの多段処理により構成される点にある。上記の群化プロセスのうちの前半のステップは、各時刻で調波構造をなしている周波数成分を見つけ出すための処理であり、隣接する時刻と独立な周波数方向の群化に相当する。そして、後半のステップは、抽出された成分あるいはそれに相当する特徴量の時系列を時間方向にスムージングする処理であり、時間方向の群化に相当する。

しかしながら、周波数方向と時間方向の群化を順次行うことが群化プロセスの最適な実践方法であるとは必ずしも言えない。なぜなら、周波数方向の群化の精度は隣接する時刻間での成分の連続性を考慮することではるかに向上されうるし、時間方向の群化の精度もまた周波数方向の群化が高精度であればあるほど高くなるはずであることから、まさに両者は、以上の意味で鶏と卵の関係にあるからである。我々は以上の問題意識のもと、より良い群化プロセスの実践のためには、個々の音源の時間周波数全域に渡ったスペクトル構造を一挙に推定できる方法論が不可欠であると考え、この観点から問題の定式化を目指した。本稿では、調波時間構造化クラスタリングと呼ぶ CASA のための新しいアプローチを提案する。

2. 調波時間構造化クラスタリングの定式化

2.1 方法論の概要と本章の構成

入力された音響信号の時間周波数成分を音脈に対応する時間周波数成分に分解する問題を定式化するにあたり、まず知覚上の単位である音脈の概念を計測可能な物理量として具現化する必要があるが、ここで言う知覚と物理量の橋渡し役を担うのがまさに Bregman の分凝要件である。そして、Bregman の分凝要件から逸脱しない範囲の自由度をもった時変スペクトルを直接的にモデル化し、これを混合したもので対象の時変スペクトルを説明しようとすることが、本稿で提案する調波時間構造化クラスタリングのアイディアの要点である。すなわち、分解された要素 (時間周波数成分) を音脈に対応する時変スペクトル成分に文字通りクラスタ化することが狙いである。

定 Q フィルタバンクが聴覚末梢系を良く模擬したモデルであることが知られることから、次節で、調波構造と共通周波数変化をなす擬似周期信号モデルの定 Q フィルタバンク出力を導出し、続く 2.3 節ではさらに共通の立上りと共通振幅変化の制約、2.4 節では成分の連続性の仮定を新たに設け、音脈に対応する時変スペクトル構造を具体的にモデル化する。2.5 節でモデルパラメータを最適化する反復アルゴリズムを導出する。

2.2 擬似周期信号の定 Q フィルタバンク出力の導出

音源 k の信号モデルとして、n 次調波成分の瞬時位相が $n\theta_k(t)+\varphi_{k,n}$、瞬時振幅が $\widetilde{w}_{k,n}(t)$ の擬似周期信号の解析信号表現

$$f_k(t)=\sum_{n=1}^{N}\widetilde{w}_{k,n}(t)e^{j\left(n\theta_k(t)+\varphi_{k,n}\right)} \tag{1}$$

を考え、この信号の定 Q フィルタバンク出力を導出する。t は時刻を表す。まず、wavelet 基底関数を

$$\psi_{a,b}(t)\triangleq\frac{1}{\sqrt{2\pi}|a|}\psi\left(\frac{t-b}{a}\right) \tag{2}$$

と定義する。ただし、a はスケールパラメータ、b はシフトパラメータ、$\psi(t)$ はアドミッシブル条件を満たす周波数 1 の任意のアナライジング wavelet である。$f_k(t)$ の連続 wavelet 変換は

$$\begin{aligned}W_k\big(\log\tfrac{1}{a},b\big)&\triangleq\Big\langle f_k(t),\psi_{a,b}(t)\Big\rangle_{t\in\mathbb{R}}\\&=\int_{-\infty}^{\infty}\sum_{n=1}^{N}\widetilde{w}_{k,n}(t)e^{j\left(n\theta_k(t)+\varphi_{k,n}\right)}\psi^*_{a,b}(t)\mathrm{d}t\end{aligned} \tag{3}$$

で与えられるが、ここで、一般に $\psi^*_{a,b}(t)$ は時刻 b に局在すること、瞬時位相 $\theta_k(t)$ と調波成分の瞬時振幅 $\widetilde{w}_{k,n}(t)$ は緩やかに変化することを前提とし、時刻 b 周辺で $\theta_k(t)$ と $\widetilde{w}_{k,n}(t)$ を

$$\widetilde{w}_{k,n}(t)\approx\widetilde{w}_{k,n}(b),\quad \theta_k(t)\approx\theta_k(b)+\theta'_k(b)\big(t-b\big) \tag{4}$$

と 0 次および 1 次近似する。瞬時位相の導関数は瞬時周波数なので、$\theta'_k(t)$ は瞬時ピッチ周波数を表し、これを $\mu_k(t)$ と置くことにすると、式 (4) より式 (3) は、

$$\begin{aligned}&W_k\big(\log\tfrac{1}{a},b\big)\approx\\&\sum_{n=1}^{N}\widetilde{w}_{k,n}(b)e^{j\left(n\theta_k(b)+\varphi_{k,n}\right)}\int_{-\infty}^{\infty}e^{jn\mu_k(b)(t-b)}\psi^*_{a,b}(t)\mathrm{d}t\end{aligned}$$

と近似できる。また、一般化 Parseval の等式より

$$\begin{aligned}&\int_{-\infty}^{\infty}e^{jn\mu_k(b)(t-b)}\psi^*_{a,b}(t)\mathrm{d}t\\&=\left\langle e^{jn\mu_k(b)(t-b)},\frac{1}{\sqrt{2\pi}|a|}\psi\left(\frac{t-b}{a}\right)\right\rangle_{t\in\mathbb{R}}\\&=\left\langle\sqrt{2\pi}\delta\big(\omega-n\mu_k(b)\big),\frac{1}{\sqrt{2\pi}}\Psi(a\omega)\right\rangle_{\omega\in\mathbb{R}}=\Psi^*\big(an\mu_k(b)\big)\end{aligned}$$

であるので、結局

$$W_k\big(\log\tfrac{1}{a},b\big)\approx\sum_{n=1}^{N}\widetilde{w}_{k,n}(b)\Psi^*\big(an\mu_k(b)\big)e^{j\left(n\theta_k(b)+\varphi_{k,n}\right)}$$

である。あとはこれを K 個分足したもの

$$W\big(\log\tfrac{1}{a},b\big)=\sum_{k=1}^{K}W_k\big(\log\tfrac{1}{a},b\big) \tag{5}$$

が混合音信号の時刻 b における定 Q フィルタバンク出力となる。ところで、a の次元は周期であるので、対数周波数 $x=\log\frac{1}{a}$

に変数変換し、$\Omega_k(b) \triangleq \log \mu_k(b)$ と置くと

$$W(x,b) = \sum_k \sum_n \widetilde{w}_{k,n}(b) \Psi^* \left(n e^{-x+\Omega_k(b)} \right) e^{j\left(n\theta_k(b)+\varphi_{k,n}\right)} \quad (6)$$

と書ける。ここで、周波数特性 $\Psi(\omega)$ が、次のような $\omega = 1$ で最大値をとる単峰的な実関数 (図 1 参照)

$$\Psi(\omega) = \Psi^*(\omega) = \begin{cases} \exp\left(-\frac{\left(\log \omega\right)^2}{4\sigma^2} \right) & (\omega > 0) \\ 0 & (\omega \leqq 0) \end{cases} \quad (7)$$

となるようなアナライジング wavelet を選ぶと、式 (6) は、

$$W(x,b) = \sum_{k,n} \widetilde{w}_{k,n}(b) \exp\left(-\frac{\left(x - \Omega_k(b) - \log n\right)^2}{4\sigma^2} \right) e^{j\left(n\theta_k(b)+\varphi_{k,n}\right)}$$

となる。以上より、式 (5) のパワースペクトルは

$$\left\| W(x,b) \right\|^2 = \sum_{k,n} \left\| \widetilde{w}_{k,n}(b) \exp\left(-\frac{\left(x - \Omega_k(b) - \log n\right)^2}{4\sigma^2} \right) e^{j\left(n\theta_k(b)+\varphi_{k,n}\right)} \right\|^2$$
$$+ \sum_{k \neq k'} \sum_{n \neq n'} \widetilde{w}_{k,n}(b) \widetilde{w}_{k',n'}(b) \exp\left(-\frac{\left(x - \Omega_k(b) - \log n\right)^2}{4\sigma^2} \right)$$
$$\exp\left(-\frac{\left(x - \Omega_{k'}(b) - \log n'\right)^2}{4\sigma^2} \right) e^{j\left(n\theta_k(b)+n'\theta_{k'}(b)+\varphi_{k,n}+\varphi_{k',n'}\right)}$$

と書けるが、ここで、上式第二項において $k \neq k'$ または $n \neq n'$ の exp 項同士の重なりがほとんどない(注1) と仮定すれば、第二項は第一項に比べると無視できるほど小さいと見なせるため、$w_{k,n}(b) \triangleq \sqrt{2\pi}\sigma \|\widetilde{w}_{k,n}(b)\|^2$ と置くと、時刻 b における混合音信号モデルの定 Q フィルタバンク出力パワーは各 Gauss 分布関数が $x = \Omega_k(b) + \log n$ でピークをとるような混合正規分布と同形の関数

$$\left\| W(x,b) \right\|^2 \approx \sum_{k,n} \frac{w_{k,n}(b)}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(x - \Omega_k(b) - \log n\right)^2}{2\sigma^2} \right) \quad (8)$$

となる (図 2 参照)。

2.3 共通振幅変化の制約

音源 k のモデルにおいて、各調波成分の共通の立上りと共通の振幅変化を仮定することは、n 次調波成分の瞬時振幅を

$$w_{k,n}(t) = \widetilde{v}_{k,n} u_k(t) \quad (9)$$

と置くことに相当する。ここで、$u_k(t)$ が

$$\int_{-\infty}^{\infty} u_k(t) \mathrm{d}t = 1 \quad (10)$$

(注1)：この仮定は、音声スパース性と呼ばれる混合音声信号の時間周波数成分はまばらにしか存在しないという性質に基づくものであり、多チャンネルブラインド音源分離においては多用される [12]。ここでは、パワースペクトルの加法性の正当化のためにスパース性の仮定を利用している。

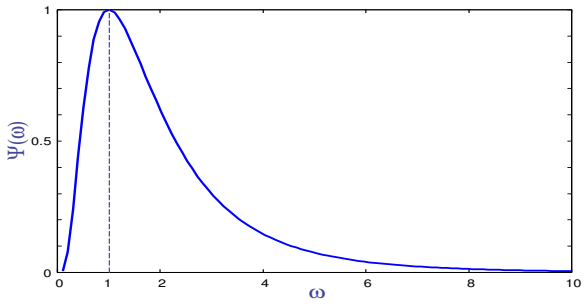

図 1　周波数特性 $\Psi(\omega)$(式 (7)) の $\sigma = \frac{1}{2}$ のときの概形。

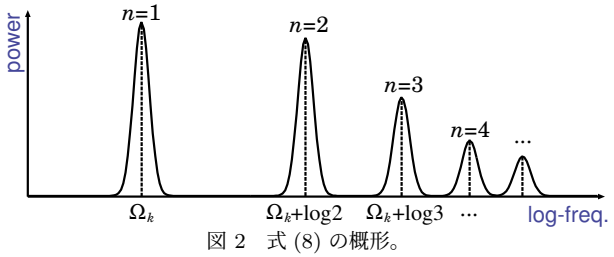

図 2　式 (8) の概形。

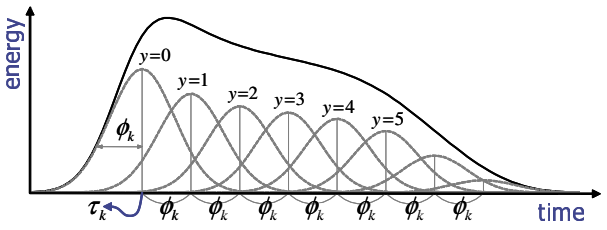

図 3　式 (11) の概形。

を満たすものとすると、パラメータ $\widetilde{v}_{k,n}$ は音源 k の n 次調波成分のエネルギー $\int_{-\infty}^{\infty} w_{k,n}(t)\mathrm{d}t$ に対応する。また、必ずしも必要ではないが、便宜的に $\widetilde{v}_{k,n} \triangleq w_k v_{k,n}$ と分解し、$v_{k,n}$ は、$\sum_n v_{k,n} = 1$ を満たすものとする。正規化共通瞬時振幅 $u_k(t)$ はノンパラメトリックに定めることもできるが、ここではパラメトリックモデルを仮定することにし、$u_k(t)$ を次のような拘束つき混合正規分布モデル (図 3 参照)

$$u_k(t) = \sum_{y=0}^{Y-1} \frac{u_{k,y}}{\sqrt{2\pi}\phi_k} \exp\left(-\frac{(t - \tau_k - y\phi_k)^2}{2{\phi_k}^2} \right) \quad (11)$$

とする。従って、式 (10) を満たすには、$\sum_y u_{k,y} = 1$ でありさえすれば良い。この選択は、非負制約や一定の滑らかさの制約を導入しやすくするだけでなく、のちに述べる効果的な最適化アルゴリズムの導出に大変都合が良い。τ_k は先頭の Gauss 分布関数の中心を表すため、これは音源 k の立上り時刻に相当するパラメータと見なせる。式 (9)、(11) を式 (8) に代入すると、具体的に

$$\sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{y=0}^{Y-1} \frac{w_k v_{k,n} u_{k,y}}{2\pi\sigma\phi_k} e^{-\frac{(x-\Omega_k(b)-\log n)^2}{2\sigma^2} - \frac{(b-\tau_k-y\phi_k)^2}{2{\phi_k}^2}} \quad (12)$$

を得る (図 4 参照)。

2.4 成分の連続性の制約

未知の連続関数であるピッチ軌跡関数 $\Omega_k(b)$ は、Weierstrass の近似定理に則り、b の多項式

$$\Omega_k(b) \triangleq \Omega_{k,0} + \Omega_{k,1} b + \Omega_{k,2} b^2 + \Omega_{k,3} b^3 + \cdots \quad (13)$$

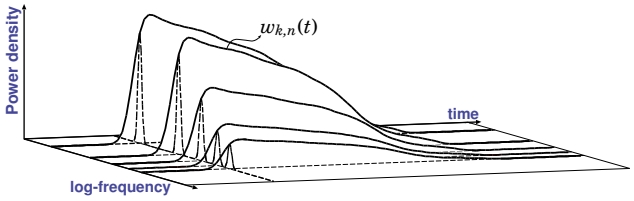

図 4　分凝要件に基づく時変パワースペクトルモデルの概形。

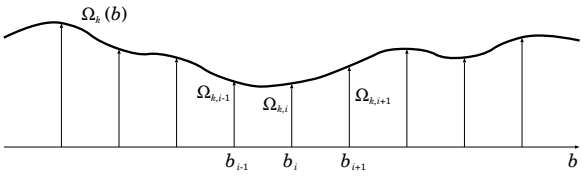

図 5　3 次 Spline ピッチ軌跡関数 (式 (14))

と、より安定な曲線を形成する区分的多項式 (3 次 Spline 関数)

$$\Omega_k(b) \triangleq \frac{1}{b_{i+1}-b_i}\Big(\Omega_{k,i}(b_{i+1}-b)+\Omega_{k,i+1}(b-b_i) - \frac{1}{6}(b-b_i)(b_{i+1}-b)\big[(b_{i+2}-b)\Omega''_{k,i}+(b-b_{i-1})\Omega''_{k,i+1}\big]\Big), \quad b\in[b_i,b_{i+1}[\quad (14)$$

(図 5 参照) の両者を検討する。ただし、上記 3 次 Spline ピッチ軌跡関数において、時間軸上のノード $b_1,\cdots,b_I$ は定数であり、各ノード上でのピッチ周波数値 $\Omega_{k,1},\cdots,\Omega_{k,I}$ がパラメータである。また、各ノード上でのピッチ軌跡関数の 2 次微分係数 $\Omega''_{k,i}$ は、$\Omega_{k,i}$ 、$\Omega''_{k,i}$ をそれぞれ要素にもつベクトルを $\boldsymbol{\Omega}$、$\boldsymbol{\Omega}''$ とすると、$\boldsymbol{\Omega}''=\boldsymbol{M}\boldsymbol{\Omega}$ で求められる。$\boldsymbol{M}$ は、$b_1,\cdots,b_I$ が定数 (既知) であれば陽に決定される定数行列である。

以上より、$\{\Omega_{k,i}\}_{i=1}^{I}$、$\{v_{k,n}\}_{n=1}^{N}$、$\{u_{k,y}\}_{y=0}^{Y-1}$、$w_k$、$\tau_k$、$\phi_k$ がすべて求まれば、$\Omega_k(b)$ および $w_{k,n}(b)$ が決まるため、式 (1) より、音源 k の信号を再構成することができる。次節より、これらのパラメータを推定するための反復アルゴリズムを導く。

2.5　パラメータ推定アルゴリズム

パラメータ $\{\{\Omega_{k,i}\}_{i=1}^{I},\{v_{k,n}\}_{n=1}^{N},\{u_{k,y}\}_{y=0}^{Y-1},w_k,\tau_k,\phi_k\}_{k=1}^{K}$ をまとめて $\boldsymbol{\Theta}$ と表記することとし、対象の信号を最も良く説明する $\boldsymbol{\Theta}$ を求めることがここでの目標である。そのためにはまず、対象とする信号の定 Q フィルタバンク出力パワーを $\|Y(x,b)\|^2$ とすると、$\|Y(x,b)\|^2$ と $\|W(x,b)\|^2$ との近さを表す尺度を導入する必要がある。$\|Y(x,b)\|^2$ と $\|W(x,b)\|^2$ はいずれも非負関数であるが、非負関数間の近さを測る尺度として I ダイバージェンスと呼ばれるものがあり、

$$I(\boldsymbol{\Theta}) \triangleq \iint\left(\left\|Y(x,b)\right\|^2\log\frac{\left\|Y(x,b)\right\|^2}{\left\|W(x,b)\right\|^2} - \left(\left\|Y(x,b)\right\|^2-\left\|W(x,b)\right\|^2\right)\right)\mathrm{d}x\mathrm{d}b \quad (15)$$

と定義される。以後これを最小化する $\boldsymbol{\Theta}$ を求めるためのアルゴリズムについて議論する。

まず、負の対数関数が凸関数であることに注目すると、

$$\sum_{k,n,y} m_{k,n,y}(x,b)=1,\ \forall k,n,y: m_{k,n,y}(x,b)\in(0,1) \quad (16)$$

を満たすような任意の重み関数 $m_{k,n}(x,b)$ を導入して

$$I(\boldsymbol{\Theta}) \leqq I^{+}(\boldsymbol{\Theta},\boldsymbol{m}) = \iint\left(\sum_{k,n,y} m_{k,n,y}(x,b)\left\|Y(x,b)\right\|^2\log\frac{m_{k,n,y}(x,b)\left\|Y(x,b)\right\|^2}{\mathcal{W}_{k,n,y}(x,b)} - \sum_{k,n,y}\left(m_{k,n,y}(x,b)\left\|Y(x,b)\right\|^2-\mathcal{W}_{k,n,y}(x,b)\right)\right)\mathrm{d}x\mathrm{d}b \quad (17)$$

のような凸不等式が成り立つ。ただし、

$$\mathcal{W}_{k,n,y}(x,b) \triangleq \frac{w_k v_{k,n} u_{k,y}}{2\pi\sigma\phi_k} e^{-\frac{(x-\Omega_k(b)-\log n)^2}{2\sigma^2}-\frac{(b-\tau_k-y\phi_k)^2}{2\phi_k^2}}$$

であるが、これは具体的には時刻 $\tau_k+y\phi_k$ 近傍にのみ分布する、n 次調波成分のパワースペクトルの一部を表す。この $\mathcal{W}_{k,n,y}(x,b)$ を、音脈を構成する「要素」に対応するものと考え、以後サブクラスタモデルと呼ぶ。さて、式 (17) を注意深く見てみると、右辺は実は $m_{k,n,y}(x,b)\|Y(x,b)\|^2$ と $\mathcal{W}_{k,n,y}(x,b)$ との間の I ダイバージェンスの総和に他ならないことに気づく。CASA やブラインド音源分離の分野ではバイナリマスクを用いて時間周波数の各成分を排他的に 1 つの音源にのみ帰属させることによる音源分離法 [12] が知られているが、これに対し、$m_{k,n,y}(x,b)$ は時間周波数成分パワー $\|Y(x,b)\|^2$ を要素に対応する成分に分解するためのいわば連続値マスクと見なせる。この不等式は、サブクラスタモデル $\mathcal{W}_{k,n,y}(x,b)$ と、連続値マスクにより分解された成分との I ダイバージェンスの総和をとったものが、全体同士の I ダイバージェンスより決して小さくなることはないことを示唆している。

この事実は、次のような反復推定アルゴリズムの導出へとつながる。まず、$I^{+}(\boldsymbol{\Theta},\boldsymbol{m})$ を任意の $\boldsymbol{\Theta}$ のもとで $m_{k,n,y}(x,b)$ に関して最小化すると、不等式より $I^{+}(\boldsymbol{\Theta},\boldsymbol{m})$ は $I(\boldsymbol{\Theta})$ より小さくなることなく等号成立状態に至る。導出は省略するが、$I^{+}(\boldsymbol{\Theta},\boldsymbol{m})$ を最小化する $m_{k,n,y}(x,b)$ は、変分法に基づき、

$$m_{k,n,y}(x,b)=\frac{\mathcal{W}_{k,n,y}(x,b)}{\sum_{k,n,y}\mathcal{W}_{k,n,y}(x,b)} \quad (18)$$

と解析的に得られる。この結果を式 (17) 右辺に代入すると、たしかに $I^{+}(\boldsymbol{\Theta},\boldsymbol{m})=I(\boldsymbol{\Theta})$ となることが分かる。続いて、この状態から $I^{+}(\boldsymbol{\Theta},\boldsymbol{m})$ を最小化する、あるいは減少させる $\boldsymbol{\Theta}$ を求めることができるならば、不等式 (17) より、減少させられた $I^{+}(\boldsymbol{\Theta},\boldsymbol{m})$ よりさらに $I(\boldsymbol{\Theta})$ は小さくなっているはずである (図 6 参照)。従って、次のような 2 ステップの反復計算

Step 0　初期値 $\boldsymbol{\Theta}_0$ を選び、$\ell=1$ とおく。
Step 1　$\boldsymbol{m}^{(\ell)}=\underset{\boldsymbol{m}}{\operatorname{argmin}}\, I^{+}(\boldsymbol{\Theta}^{(\ell-1)},\boldsymbol{m})$
Step 2　$I^{+}(\boldsymbol{\Theta},\boldsymbol{m}^{(\ell)})\leqq I^{+}(\boldsymbol{\Theta}^{(\ell-1)},\boldsymbol{m}^{(\ell)})$ なる $\boldsymbol{\Theta}$ を $\boldsymbol{\Theta}^{(\ell)}$、$\ell\leftarrow\ell+1$ として Step 1 へ戻る。

を行うことで $I(\boldsymbol{\Theta})$ を単調減少させていくことができる。$I(\boldsymbol{\Theta})$ は下に有界なので、上記の反復計算は収束性が保証される。

この反復計算は、$I(\boldsymbol{\Theta})$ の最小化が解析的に行えない代わりに $I^{+}(\boldsymbol{\Theta},\boldsymbol{m})$ を最小化あるいは減少させられる $\boldsymbol{\Theta}$ の更新式が解析的に得られる場合に特に効果的である。ところで、この反復計算は、実は、EM(Expectation-Maximization) アルゴリズム [13] と形式上は等価である。しかしながら、文献 [13] の導出に直接従えば、$\|Y(x,b)\|^2$ と $\|W(x,b)\|^2$ がいずれも確率密度関数でなければならず、パワースペクトルに対して Bayes 則

(条件つき確率の公式や周辺化操作) が適用できるかどうかは自明ではない。その意味で、上記のように Bayes 則を適用せずとも EM アルゴリズムと形式的に等価な反復アルゴリズムが導けることは、本問題設定上においては重要な意味をもつ。

Step 2 における $\{\Omega_{k,i}\}_{i=1}^{I}$,$\{v_{k,n}\}_{n=1}^{N}$,$\{u_{k,y}\}_{y=0}^{Y-1}$,$w_k$,$\tau_k$,$\phi_k$ の更新式は、導出は省略するが、

$$
\begin{aligned}
w_k^{(\ell)} &= \sum_{n,y} \iint \Phi_{k,n,y}^{(\ell)}(x,b)\mathrm{d}x\mathrm{d}b \\
\Omega_{k,0}^{(\ell)} &= \frac{1}{w_k^{(\ell)}} \sum_{n,y} \iint \Phi_{k,n,y}^{(\ell)}(x,b)(x-\log n)\mathrm{d}x\mathrm{d}b \\
\tau_k^{(\ell)} &= \frac{1}{w_k^{(\ell)}} \sum_{n,y} \iint \Phi_{k,n,y}^{(\ell)}(x,b)(t-y\phi_k^{(\ell-1)})\mathrm{d}x\mathrm{d}b \\
v_{k,n}^{(\ell)} &= \frac{1}{w_k^{(\ell)}} \sum_{y} \iint \Phi_{k,n,y}^{(\ell)}(x,b)\mathrm{d}x\mathrm{d}b \\
u_{k,y}^{(\ell)} &= \frac{1}{w_k^{(\ell)}} \sum_{n} \iint \Phi_{k,n,y}^{(\ell)}(x,b) \\
\phi_k^{(\ell)} &= \frac{1}{2w_k^{(\ell)}} \left(\left(\alpha_k^{(\ell)2} + 4\beta_k^{(\ell)} w_k^{(\ell)} \right)^{\frac{1}{2}} - \alpha_k^{(\ell)} \right) \\
&\begin{cases}
\alpha_k^{(\ell)} \triangleq \sum_{n,y} \iint \Phi_{k,n,y}^{(\ell)}(x,b) y(t-\tau_k^{(\ell)})\mathrm{d}x\mathrm{d}b \\
\beta_k^{(\ell)} \triangleq \sum_{n,y} \iint \Phi_{k,n,y}^{(\ell)}(x,b)(t-\tau_k^{(\ell)})^2\mathrm{d}x\mathrm{d}b
\end{cases}
\end{aligned}
$$

のように解析的に得られる。ただし、

$$
\Phi_{k,n,y}^{(\ell)}(x,b) = m_{k,n,y}^{(i)}(x,b) \left\| Y(x,b) \right\|^2
$$

である。なお、上記の $\Omega_{k,0}$ は、ピッチ周波数軌跡が多項式 (式 (13)) の場合の 0 次項の係数である。1 次以上の項の係数の更新式も同様に解析的に得られるが、ここでは紙面の都合上省略する。一方で、ピッチ周波数軌跡が 3 次 Spline 関数 (式 (14)) の場合の各係数の更新式は、

$$
\Omega_{k,i}^{(\ell)} = \frac{\displaystyle\sum_{n,y} \iint \left(x - \widetilde{\Omega}_{k,n,i}^{(\ell)}\left(b; \boldsymbol{\Omega}_{k,i}^{(\ell)}\right) \right) \frac{\partial \Omega_k(b)}{\partial \Omega_{k,i}} \Phi_{k,n,y}^{(\ell)}(x,b)\mathrm{d}x\mathrm{d}b}{\displaystyle\sum_{n,y} \iint \left(\frac{\partial \Omega_k(b)}{\partial \Omega_{k,i}} \right)^2 \Phi_{k,n,y}^{(\ell)}(x,b)\mathrm{d}x\mathrm{d}b}
$$

で与えられる。ただし、$\boldsymbol{\Omega}_{k,i}^{(\ell)} = (\Omega_{k,1}^{(\ell)}, \ldots, \Omega_{k,i-1}^{(\ell)}, \Omega_{k,i}^{(\ell-1)}, \Omega_{k,i+1}^{(\ell-1)}, \ldots, \Omega_{k,I}^{(\ell-1)})$ とする。また、$\widetilde{\Omega}_{k,n,i}^{(\ell)}(b; \boldsymbol{\Omega}_{k,i}^{(\ell)}) = \Omega_k(b; \boldsymbol{\Omega}_{k,i}^{(\ell)}) - \frac{\partial \Omega_k(b)}{\partial \Omega_{k,i}} \Omega_{k,i}^{(\ell)} + \log n$ であり、この項は $\Omega_{k,i}$ に依らない。$\frac{\partial \Omega_k(b)}{\partial \Omega_{k,i}}$ は b と定数行列 $\boldsymbol{M}$ にのみ依る項である。ところで、σ は定 Q フィルタバンクの解析条件で決まる定数であったが、これも音源 k ごとの変数と見なして推定することもできる。その場合、σ_k の更新式は同様にして

$$
\sigma_k^{(\ell)} = \left(\frac{1}{w_k^{(\ell)}} \sum_{n,y} \iint \Phi_{k,n,y}^{(\ell)}(x,b)(x - \Omega_k^{(\ell)}(b) - \log n)^2 \mathrm{d}x\mathrm{d}b \right)^{\frac{1}{2}}
$$

と求まる。

また、本稿では紙面の都合上割愛せざるを得なかったが、式 (15) の最小化は、回帰分析の問題と捉えると、$\|W(x,b)\|^2$ をパラメータとする $\|Y(x,b)\|^2$ の連続 Poisson 分布の結合分布を尤度とした最尤推定と等価であることが分かっており、それにより、適当なパラメータの事前分布を定めることにより事後確率推定最大パラメータを同等な反復アルゴリズムで求めることができる。これにより、他のさまざまな先験的制約をスムーズに導入できる点もここで強調しておく (文献 [15], [16] 参照)。

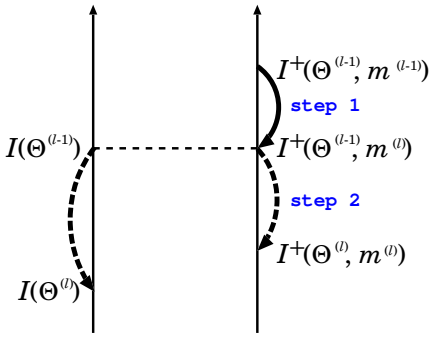

図 6　パラメータ反復推定アルゴリズムの概念図。

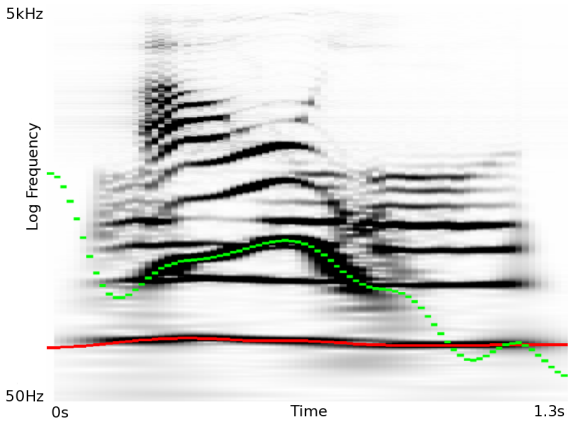

図 7　混合音声信号のスペクトログラムと推定されたピッチ軌跡関数

3.　性能の定量的評価

聴覚においてひとまとりと知覚される単位が物理的な意味での一連の音響信号と対応するとは限らないが、実際のアプリケーションを想定して提案する方法論がどの程度の性能を示すかを評価実験により確認することは、方法論自体の工学的な実用可能性だけでなく、Bregman の分凝要件がいかに物理的事実と結びついているかを示す 1 つの目安にもなると考えられる。

3.1　混合音声のピッチ周波数推定精度

まず、提案法の動作確認のため、単一話者音声のピッチ周波数の推定精度を評価した。女性と男性話者による計 100 個の英文の読み上げ音声の Laryngograph 信号 [14] を実験データとした。ここでは、HTC のピッチ周波数軌跡は 3 次スプライン関数を仮定することとした。分析条件やピッチ周波数の正解データの作成方法などに関する詳細は文献 [16] に示されている。推定精度は、サンプリング周期ごとに推定ピッチ周波数と正解ピッチ周波数を照らし合せ、正解からの誤差が 20%以内だったものを正解として算出した。表 1 に、現在最先端のピッチ周波数推定手法である YIN [17] やそれ以外の手法との誤り率の比較を示す。YIN 以外の略称が指す手法名については文献 [16] に詳しく示されている。単ピッチ推定に特化したこれらの従来手法と HTC の性能が遜色がないことは特筆すべき結果と言える。

続いて、同じ分析条件で、2 話者 (男性と女性、男性同士、女性同士) の音声を平均パワーが等しくなるように混合した 150 個の混合音声信号を対象とし、混合音声信号のピッチ推定精度を評価した。CASA の最先端の手法の 1 つである Wu らの手

表 1　単ピッチ推定精度の種々の従来手法との比較

Method	pda	fxac	fxcep	ac	cc	shs	acf	nacf	additive	TEMPO	YIN	HTC
Gross error (%)	19.0	16.8	15.8	9.2	6.8	12.8	1.9	1.7	3.6	3.2	1.4	3.5

表 2　混合音声信号のピッチ推定精度の WWB 法との比較

Gross error threshold	20%		10%	
methods	HTC	WWB	HTC	WWB
男性と女性	93.3	81.8	86.8	81.5
男性同士	96.1	83.4	87.9	69.0
女性同士	98.9	95.8	95.6	90.8
Total	96.1	87.0	90.2	83.5

法 (以後、WWB 法) [11] との性能比較を表 2 に示す。ここでは、正解と見なす許容誤差を ±10% と ±20% とした場合の両者の精度を評価し、HTC がいずれにおいても WWB 法を上回る成績を得た。提案法の動作を直感的に確認するため、混合音声信号のスペクトログラムと推定されたピッチ軌跡関数を重ねて表示したものを図 7 に示す。

3. 2　音楽音響信号の音高推定精度

次に、音楽音響信号を対象とし、音名の推定精度を評価するため、RWC 研究用音楽データベース [18] よりクラシックとジャズの 8 楽曲を実験データに選び、PreFEst [19] と呼ぶ最先端の音楽の音高推定手法と性能比較を行った。ここでの実装では、HTC のピッチ軌跡関数は 0 次多項式とした。なお、その他の分析条件や正解データの作成方法などに関する詳細は文献 [15] に示されている。実験の結果、PreFEst は平均音名推定精度が 62.4%であったのに対し、HTC は 70.4%という成績を得、音楽の音高推定においても HTC が効果的であることを確認した。時変スペクトルモデルがどの程度正確に推定されているかを視覚的に確認するため、対象とする信号のスペクトログラムと並べて最適推定された時変スペクトルモデルを図 8 に示す。

4. ま と め

本稿では、Bregman の分凝要件に立脚し、個々の音源の時間周波数全域に渡ったスペクトル構造を一挙に推定できる新しい方法論を提案し、これを調波時間構造化クラスタリング (HTC) と名づけた。混合音声信号および音楽音響信号のピッチ周波数推定精度の評価実験を通し、それぞれの分野における最先端の従来法の精度を上回る結果を得た。

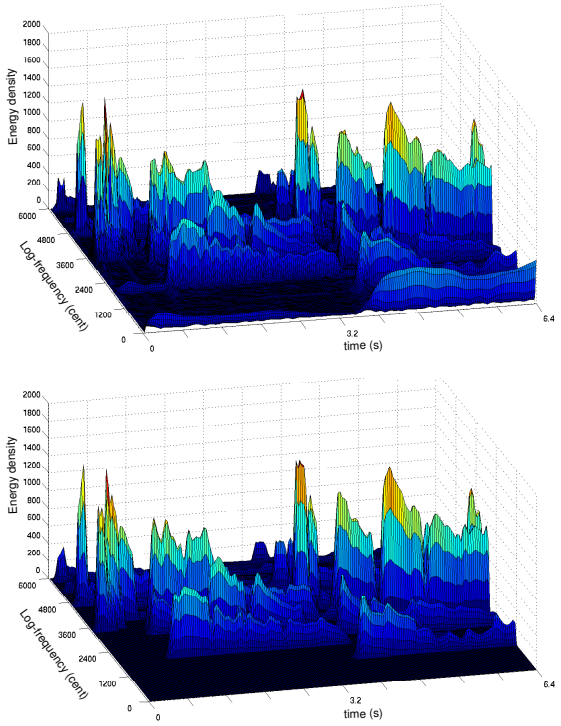

図 8　対象のスペクトログラム (上) と推定された時変スペクトルモデル (下)

文　献

[1] A. S. Bregman, Auditory Scene Analysis, MIT Press, Cambridge, 1990.

[2] M. Cooke, G. J. Brown, M. Crawford and P. Green, “Computational Auditory Scene Analysis – Listening to Several Things at Once,” Endevour New Series, Vol. 17, No. 4, pp. 186–190, 1993.

[3] G. J. Brown and M. Cooke, “Computational Auditory Scene Analysis,” Comp. Speech & Lang., No. 8, pp. 297–336, 1994.

[4] D. P. W. Ellis, “A Computer Implementation of Psychoacoustic Grouping Rules,” In Proc. IEEE ICPR’94, pp. 108–112, 1994.

[5] A. Fishbach, “Primary Segmentation of Auditory Scene,” In Proc. IEEE ICPR’94, pp. 113–117, 1994.

[6] T. Nakatani, M. Goto and H. G. Okuno, “Localization by Harmonic Structure and Its Application to Harmonic Sound Segregation,” In Proc. IEEE ICASSP’96, pp. 653–656, 1996.

[7] 西, 安部, 安藤, “聴覚情景分析のための多重ピッチ追跡と調波分離アルゴリズム,” 計測自動制御学会, Vol. 34, No. 6, pp. 483–490, 1998.

[8] 鵜木, 赤木, “聴覚の情景分析に基づいた雑音下の調波複合音の一抽出法,” 電子情報通信学会論文誌, Vol. J82-A, No. 10, pp. 1497–1507, 1999.

[9] 安部, 安藤, “共有 FM-AM の時間周波数統合に基づく聴覚情景分析 (I)—Lagrange 微分特徴量とその周波数統合—,” 電子情報通信学会論文誌, Vol. J83-D-II, No. 2, pp. 458–467, 2000.

[10] 安部, 安藤, “共有 FM-AM の時間周波数統合に基づく聴覚情景分析 (II)—最適な時間軸統合とストリーム音の再合成—,” 電子情報通信学会論文誌, Vol. J83-D-II, No. 2, pp. 468–477, 2000.

[11] M. Wu, D. L. Wang and G. J. Brown, “A Multipitch Tracking Algorithm for Noisy Speech,” IEEE Trans., Speech and Audio Process., Vol. 11, pp. 229–241, 2003.

[12] Ö. Yilmaz and S. Rickard, “Blind Separation of Speech Mixtures via Time-Frequency Masking,” IEEE Trans., Signal Process., Vol. 52, No. 7, pp. 1830–1847, 2004.

[13] A. P. Dempster, N. M. Laird and D. B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” J. of Royal Statistical Society Series B, Vol. 39, pp. 1–38, 1977.

[14] P. C. Bagshaw, S. M. Hiller and M. A. Jack, “Enhanced Pitch Tracking and the Processing of F_0 Contours for Computer and Intonation Teaching,” In Proc. Eurospeech’93, pp. 1003–1006, 1993.

[15] H. Kameoka, T. Nishimoto and S. Sagayama, “A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering,” IEEE Trans., Audio, Speech, Language Process., in Press, 2006.

[16] J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigné and S. Sagayama, “Single and Multiple Pitch Contour Estimation through Parametric Spectrogram Modeling of Speech in Noisy Environments,” IEEE Trans., Audio, Speech, Language Process., submitted in 2006.

[17] A. de Cheveigné and H. Kawahara, “YIN, a Fundamental Frequency Estimator for Speech and Music,” J. Acoust. Soc. Am., 111(4), pp. 1917–1930, 2002.

[18] 後藤, 橋口, 西村, 岡, “RWC 研究用音楽データベース: 研究目的で利用可能な著作権処理済み楽曲・楽器音データベース,” 情報処理学会論文誌, Vol.45, No.3, pp.728–738, 2004.

[19] M. Goto, “A Real-Time Music-Scene-Description System: Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-World Audio Signals,” ISCA Journal, Vol. 43, No. 4, pp. 311–329, 2004.