January 5, 1999
RT0295
Computer Science 11 pages

Research Report

A study of computational auditory scene analysis

Masaharu Sakamoto,Michio Yamada(University of Tokyo)

IBM Research, Tokyo Research Laboratory
IBM Japan, Ltd.
1623-14 Shimotsuruma, Yamato
Kanagawa 242-8502, Japan

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

混合音からの音源分離

阪本 正治
日本アイ・ビー・エム (株）東京基礎研究所/東京大学数理科学研究科
山田道夫
東京大学数理科学研究科

1998 年 12 月 29 日

1 はじめに

実環境においては、複数の音源から放射された音が混ざりあって存在するのが普通である。そのような環境下でも、地下鉄の電車の騒音の中で会話ができたり、大勢の人たちの会話が飛び交う中で特定の相手の声だけに注意を向けることができるなど、人間の音声認識機能は安定に動作する。人間が持つこの能力はカクテルパーティー効果と呼ばれ、実環境下でも自動音声認識システムを安定に稼動させのに役立っている。一方、現状の自動音声認識システムは、カクテルパーティー効果に相当する機能を持っていないため、実環境下において十分な性能を発揮できない。

Bregman は、カクテルパーティー効果に代表される人間の様々な聴覚現象を、音を通じて環境を把握するための機能の現われとして捉え直すことによって統一的に理解しようとする「聴覚による情景分析」(Auditory Scene Analysis、以下 ASA と記す。) という考え方 [1] を提唱している。

この考え方自体が、実環境下での自動音声認識性能の向上に直接的に役立つわけではないが、この考え方に触発され、聴覚の情景分析機能の工学的な実現を目指して、計算機による聴覚の情景分析（Computational Auditory Scene Analysis）と呼ばれる研究分野が形成されつつある。CASA における代表的な問題は、カクテルパーティー効果の計算機による実現であり、音源分離あるいはストリーム（音脈）分離と呼ばれる。

我々は、CASA の代表的な例として、英国 Sheffield 大の Guy Brown らの研究 [2] を取り上げ、その手法を、MATLAB 言語[1]を使って実現した。

2 聴覚による情景分析 (Auditory Scene Analysis)

ASA では、聴覚を鼓膜に到達した音から周囲の状況を把握するための機能として捉える（電車が通り過ぎた、ピアノを弾いている、子供が話している、など）。一般に、鼓膜に到達する音は、複数の音源からの音が混ざり合った混合音として到達するので、鼓膜に到達した混合音から、周囲で起こったことを把握するには、脳で混合音を個々の音源の音に分離することが必要になる。ASA においては、音響信号は知覚上の要素に分けられ、次の段階で同じ音源から生じていると思われる成分どうしが統合され音源の分離が達成されると考える。

このような過程は、混合音から音源を分離するという一種の逆問題と考えることができるが、音源の数が未知であるので、そのままでは一意に解を求めることはできない。Bregman は、聴覚がこの逆問題を解くために利用している拘束条件として、以下のような規則 [11] をまとめている。

- 異なる音源からの音が同時に始まったり終わったりしない。
- 変化は急激には起こらない
- ものが繰り返し振動するときには、共通の基本周波数の整数倍の音響的成分が発生する。

[1]MATLAB は The MathWorks,Inc. の登録商標です。

- 一つの音響的事象に生ずる多くの変化は、その音を構成する各成分に同時に同じような影響を与える。

3 Guy Brownらのシステム

英国 Sheffield 大の Guy Brown らは、混合音を時間－周波数平面上での知覚上の要素に分解し、Bregman の拘束条件を利用して音響エレメントをグループ化する手法 [2] で、音源分離システムを構築している。この手法は、従来の音源分離の手法にみられる音源の数や性質（相関、ピッチレンジ）に関する拘束がなく、柔軟なシステムである。この手法のダイアグラムを図 1 に示す。

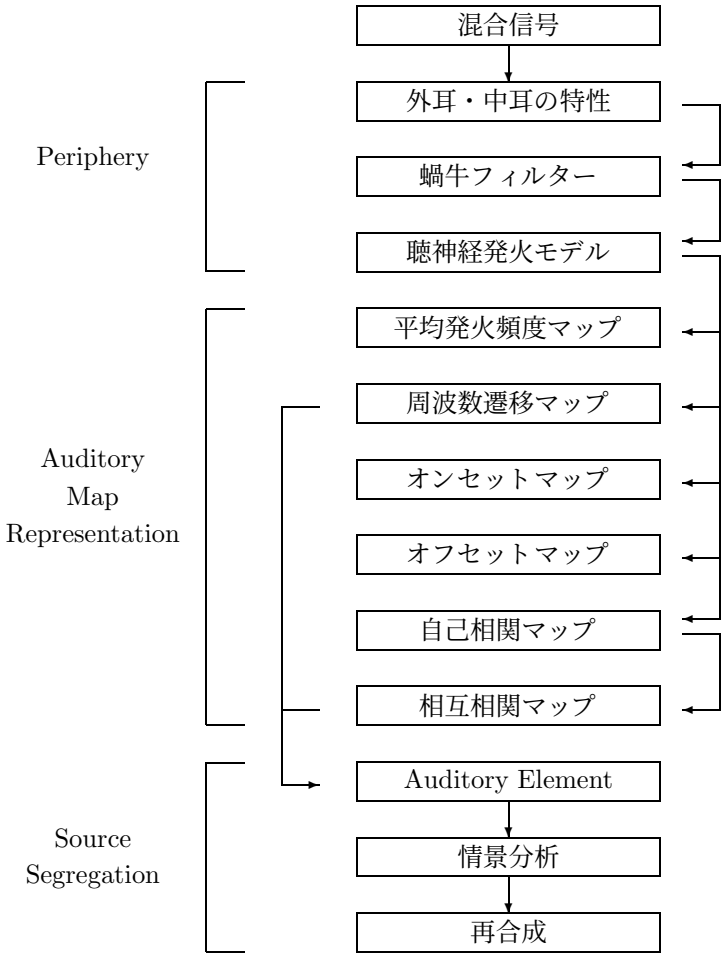

図 1: 音源分離手法の概要 [2]

3.1 聴覚末梢系（Auditory periphery model)

最初のステージでは、混合音は、外耳・中耳の特性を模擬した高域強調フィルター、蝸牛フィルター、聴神経発火モデルからなる聴覚末梢系のシミュレーションを通り、周波数チャネル毎の神経発火頻度の時系列データに変換される。[2]

[2]ここでは、耳介から蝸牛の内有毛細胞とそれに接続する 1 次聴神経までを聴覚末梢系と呼ぶ。

3.1.1 外耳と中耳

耳介と外耳道からなる外耳は音波を鼓膜に伝える音響系であり、中耳は鼓膜の振動を内耳に伝える機械系である。外耳道の音響特性は 3kHz 付近に共振周波数を持っている。一方、中耳の伝達特性は、複雑である。猫の鼓膜に対して行った中耳の周波数伝達特性の実測値によれば、700Hz 以下の周波数では伝達特性はほぼ一定であり、1000Hz 以上では約 12dB/Oct の減数特性を示し、3000Hz 以上では、不規則な特性となっている。ここでは、外耳中耳を単に高域通過フィルターとみなす。

$$x(t) = x(t) - 0.95x(t-1) \tag{1}$$

ここで、x(t) は時刻 t における入力であり、y(t) は、フィルター出力を表している。

3.1.2 蝸牛フィルター

基底膜での周波数選択性は、基底膜上の各位置での周波数応答を模擬したフィルターバンクでモデル化できる。ここで用いるフィルターバンクは、生理学的な実験データを元にして、de Boer と de Jongh によって提案されたガンマートーンフィルター [4] である。中心周波数 f_0[Hz] における n 次のガンマートーンフィルターのインパルス応答は以下の式で表される。

$$gt(t) = t^{n-1}\exp(-2\pi bt)\cos(2\pi f_0 t + \varphi) \tag{2}$$

Brown らは、ERB スケール [7] 上で等間隔に 128 チャンネルのガンマートーンフィルターを用いている。各チャネルの中心周波数は、特徴周波数と呼ばれる。

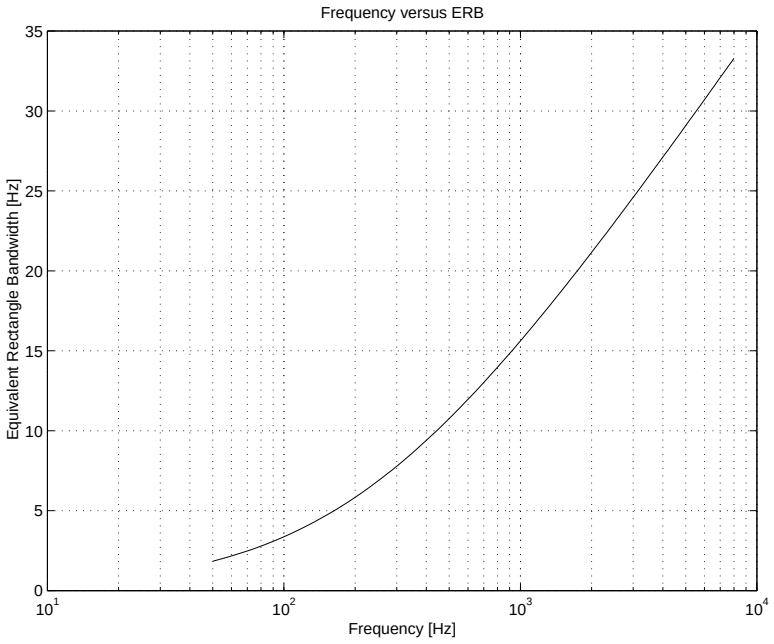

図 2: 中心周波数と等価矩形幅 (ERB) の関係

gammartone フィルターの周波数特性は、中心周波数に対してほぼ対称で、心理物理的に測定された中程度の音圧レベルの場合のヒトの聴覚フィルターの形状に近いが、周波数特性のレベル依存性が表現できない。基底膜振動のレベル依存性を模擬したモデルとして、いくつか提案されている [5][6]。

我々は、Slany[10] による IIR 型 gammatone フィルターを用いた。

3.1.3 有毛細胞シナプスモデル

Meddis による有毛細胞と聴神経シナプスでの神経伝達物質の生成、消失、移動を差分方程式で記述した機械・神経系変換の計算モデル [8][9] である。1 チャネルの蝸牛フィルターの出力に対し、1 つの有毛細胞シナプスモデルを接続することで、聴神経の鋭い周波数選択性（特徴周波数）を模擬している。蝸牛フィルター出力を入力とし、シナプス間隙に到達する神経伝達物質の量に比例したスパイク頻度を出力として取り出す。

3.2 聴覚マップ (Auditory periphery model)

第 2 のステージでは、音源の聴覚上の特徴が、様々な角度から聴覚マップによって表現される。それらのマップは、ニューロンの特徴周波数とパラメーター値の 2 次元表示となっている。Brown らは、平均発火頻度マップ、周波数遷移マップ、オンセットマップ、オフセットマップ、自己相関マップ、相互相関マップを作成している。

3.2.1 平均発火頻度マップ

Meddis の有毛細胞シナプスモデルの出力を、20ms の Hamming 窓で平滑化した平均発火頻度のマップである。

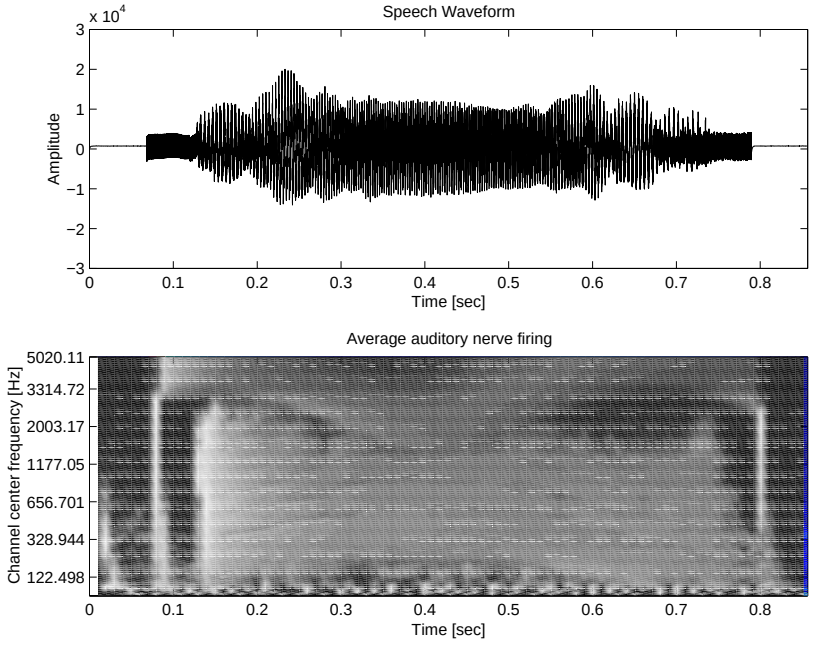

図 3: 女性による日本語発声青々」とサイレン音の混合音波形 (サンプリング周波数 22.05kHz) と平均発火頻度マップ

図 3 は、女性による「青々」という発声と、サイレン音との混合音 (サンプリング周波数 22.05kHz) の平均発火頻度マップである。スペクトログラムのように音声信号のフォルマントやサイレン音の周波数変化が観察できる。

3.2.2 周波数遷移マップ

平均発火頻度マップ上のスペクトルピークが、次の時刻にどこに移動しているかという情報を表している。周波数遷移の方向は、平均発火レート $r(t,f)$ とスペクトルピークの移動方向を抽出する関数 g_θ(receptive field) との畳み込み（式 (3)）

$$s(t,f,\theta) \quad = \quad \sum_{i=-N}^{N} \sum_{j=-M}^{M} r(t+i, f+j) g_\theta(i,j) \tag{3}$$

から算出される。すなわち、式 (5) によって、平均発火頻度マップ上でピークを持つ周波数が求まり、式 (5) によって、そのピークの遷移方向が求まる。

$$\frac{\partial}{\partial f}s(t,f,0) = 0 \tag{4}$$

$$\frac{\partial}{\partial \theta}s(t,f,\theta) = 0 \tag{5}$$

g_θ は、時間方向と周波数方向の 2 次元ガウス関数

$$G(t,f) = exp\Big(\frac{t^2}{2\pi{\sigma_t}^2} - \frac{f^2}{2\pi{\sigma_f}^2}\Big) \tag{6}$$

の 2 階微分

$$g(t,f) = \frac{\partial^2}{\partial f^2}G(t,f) \tag{7}$$

に対して、回転を施したものとして与えられる。

$$g_\theta(t,f) = gR_\theta(t,f) \tag{8}$$

ここで、

$$R_\theta(t,f) = \begin{pmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{pmatrix}\begin{pmatrix} t \\ f \end{pmatrix} \tag{9}$$

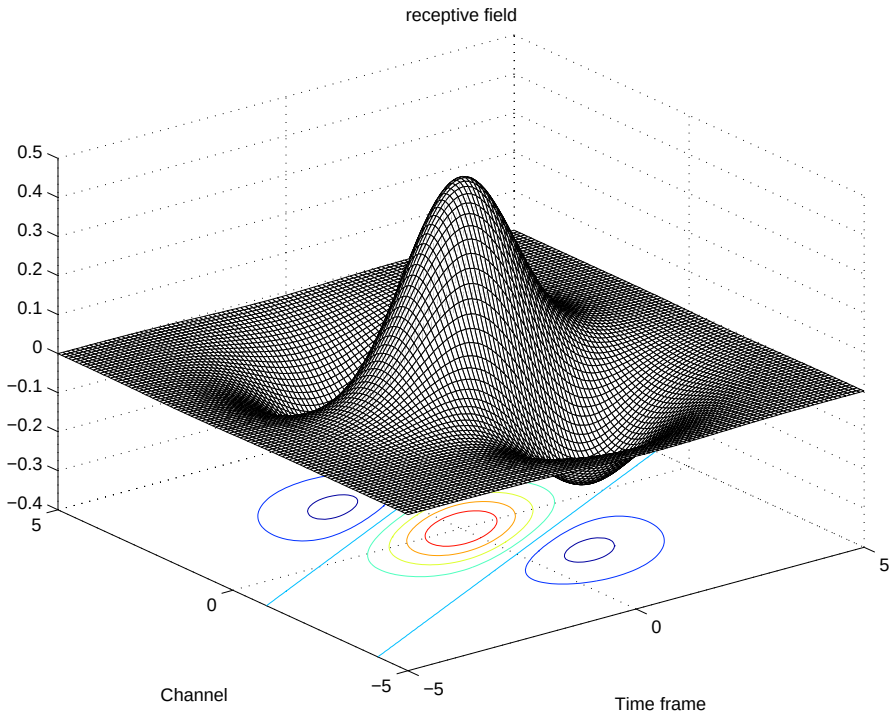

図 4: Receptive field

図 4 に、上昇方向の周波数遷移を捉えるための receptive field を示す。

図 5 は、図 3 の平均発火頻度マップに対する周波数遷移マップである。フォルマントやサイレン音の周波数変化が観察できる。

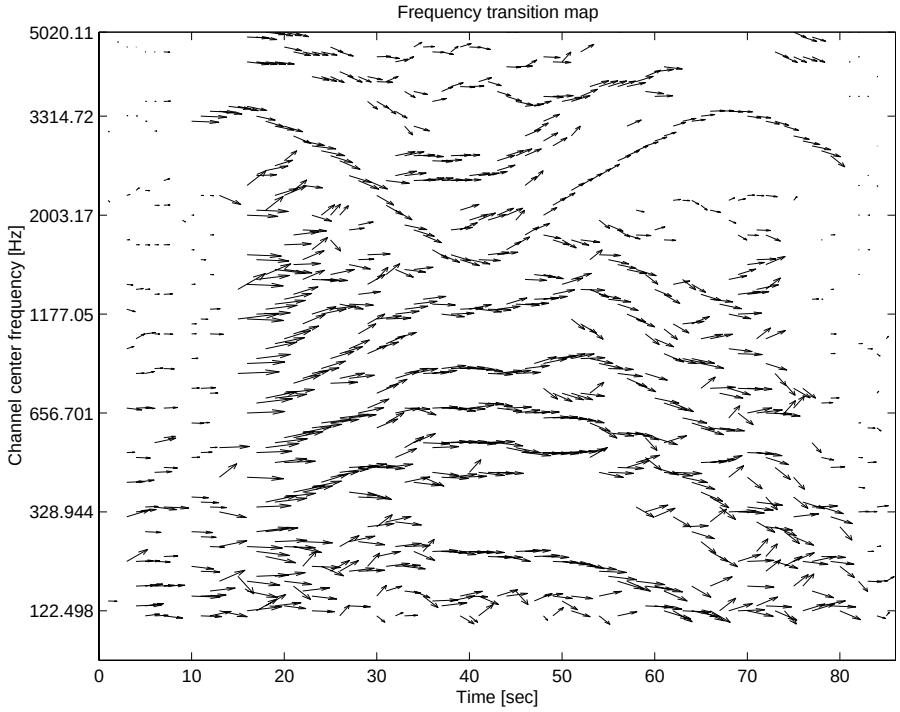

図 5: 周波数遷移マップ

3.2.3 オンセットマップ、オフセットマップ

ASA の仮定にもあるように、異なる音源からの音が、同時に始まったり、終わったりすることはない。したがって、音の始まりと終わりを拠り所に音の構成要素をグルーピングすることは合理的である。実際、聴覚中枢で音の始まりや終わりに反応する神経細胞が見つかっている。オンセットマップでは、式 (10) に示すように音刺激による興奮性の反応の直後に強い抑制性の反応が起こるようなメカニズムで実現している。

$$p_{on}(t) = p_{on}(t-1)c_d + E_{psp}r(t) - I_{psp}r(t-\Delta t_I) \tag{10}$$

$$c_d = exp\Big(-\frac{dt}{\tau_d}\Big) \tag{11}$$

$$S_{on}(t) = \begin{cases} p_{on}(t) & \text{if } p_{on}(t) > Th \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

オフセットマップは、式 (13) に示すように、逆のメカニズムで実現されている。

$$p_{off}(t) = p_{off}(t-1)c_d + E_{psp}r(t-\Delta t_E) - I_{psp}r(t) \tag{13}$$

$$S_{off}(t) = \begin{cases} p_{off}(t) & \text{if } p_{off}(t) > Th \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

図 6 は、図 3 の平均発火頻度マップに対するオンセットマップとオフセットマップである。

3.2.4 自己相関マップと相互相関マップ

このマップは、聴覚でのピッチ知覚が周波数領域でのスペクトル分析と時間領域の自己相関分析が同時に行われて達成しているという考えに基づいている。時刻 t における特徴周波数 f での自己相関 c は平均発火頻度マップ $r(t,f)$ から式 (15) で求められる。

$$c(t,f,\Delta t) = \sum_{\infty}^{i=0} r(t-T,f)r(t-T-\Delta t,f)h(T) \tag{15}$$

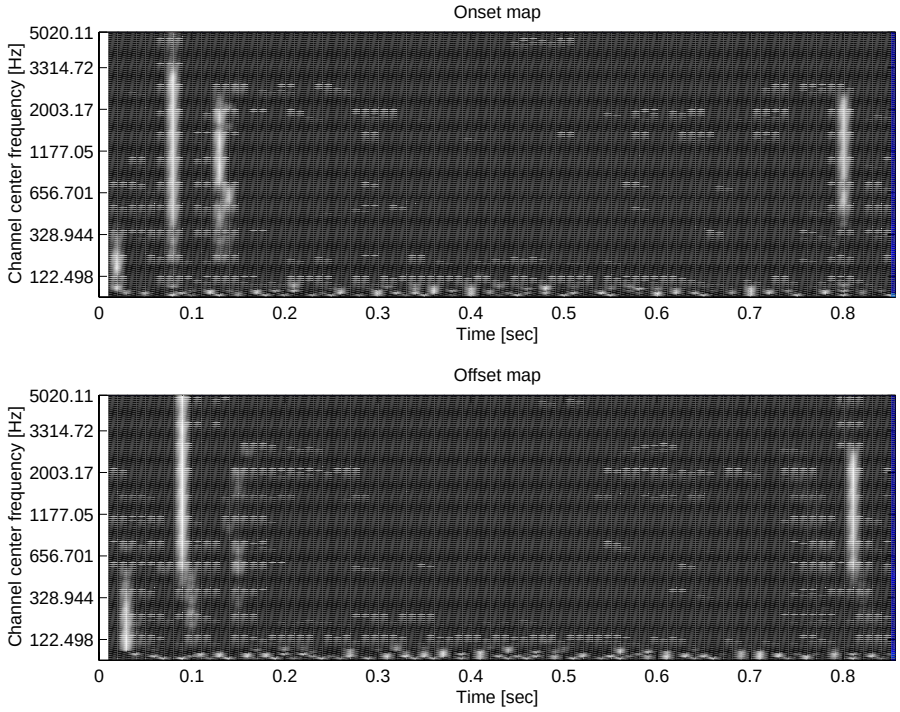

図 6: オンセットマップ、オフセットマップ

ここで、

$$\mathrm{T} = idt \tag{16}$$

自己相関マップは、次式で平均発火頻度の影響を除いたものとして、式 (17) で求められる。

$$a_n(t, f, \Delta t) \quad = \quad \frac{c(t, f, 0)}{c(t, f, \Delta t)} \tag{17}$$

自己相関マップは, 周期信号に対しては、その調波構造のため隣接するチャネルで同じようなパターンとなり冗長である。そこで、自己相関マップから式 (18) でチャネル間の相互相関を計算する。これを相互相関マップと呼ぶ。

$$sim(f_1, f_2, t) \quad = \quad \frac{2\sum_{\Delta t} a_n(t, f_1, \Delta t) a_n(t, f_2, \Delta t)}{\sum_{\Delta t} a_n(t, f_1, \Delta t)^2 + \sum_{\Delta t} a_n(t, f_2, \Delta t)^2} \tag{18}$$

相互相関マップ上で、大きい相関係数をもつチャネルどうしは、ASA の仮定にもあるように、同一音源からの成分である可能性が高い。そこで、あらかじめ定めた閾値を超えているチャンネルどうしをグルーピングする。グルーピングされたチャンネルを periodicity group と呼ぶ。しかし、単に、閾値だけでは、幾通りかのグルーピングが考えられるので、一つに決めるため、”area stability criterion”を用いる。”area stability criterion”とは、相互相関マップ上で、自分より大きい面積のグループが派生していない periodicity group を選択する方法である。図 7 は、図 3 の平均発火頻度マップの 0.33 秒付近での自己相関マップ、相互相関マップをである。相互相関マップ上の灰色のブロックが periodicity group である。図 8 は、図 3 の平均発火頻度マップ求めた全時間にわたる priodicity group である。音声の調波構造や、サイレン音がそれぞれ 1 つの periodicity group を形成している。

3.3 情景分析

最後のステージで音源分離が行われる。まず、periodicity group と周波数遷移マップから auditory elelement を形成する。各 auditory elelement 毎に、F0（基本周波数)）軌跡を求め、同じ F0 変化パターン、あるいは同じオフセット時間あるいはオンセット時間を持つ auditory elelement を探し出し、それらを統合していくことによって音源分離が行われる。

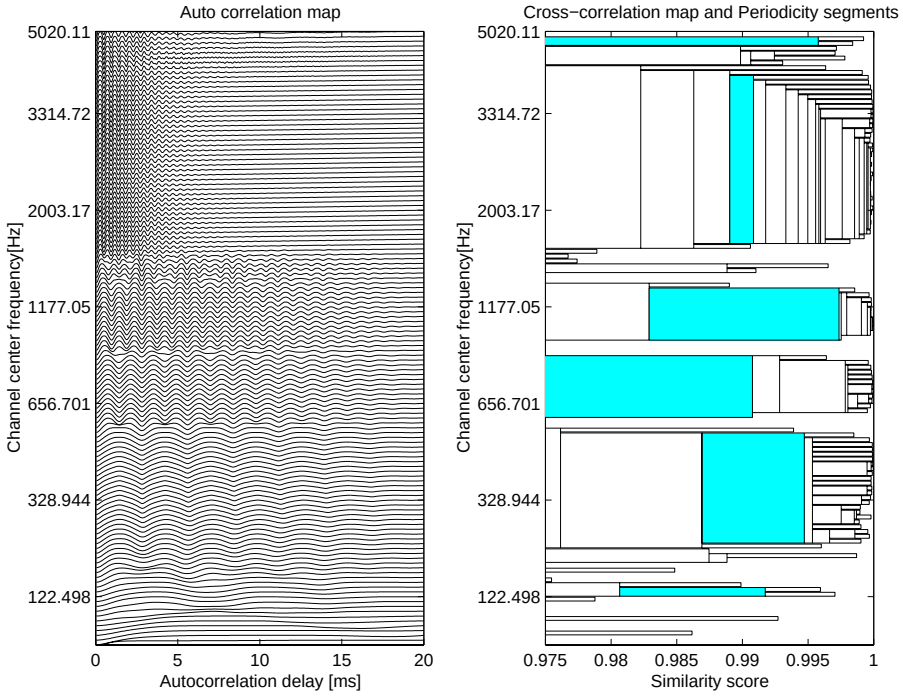

図 7: 自己相関マップ、相互相関マップと periodicity group

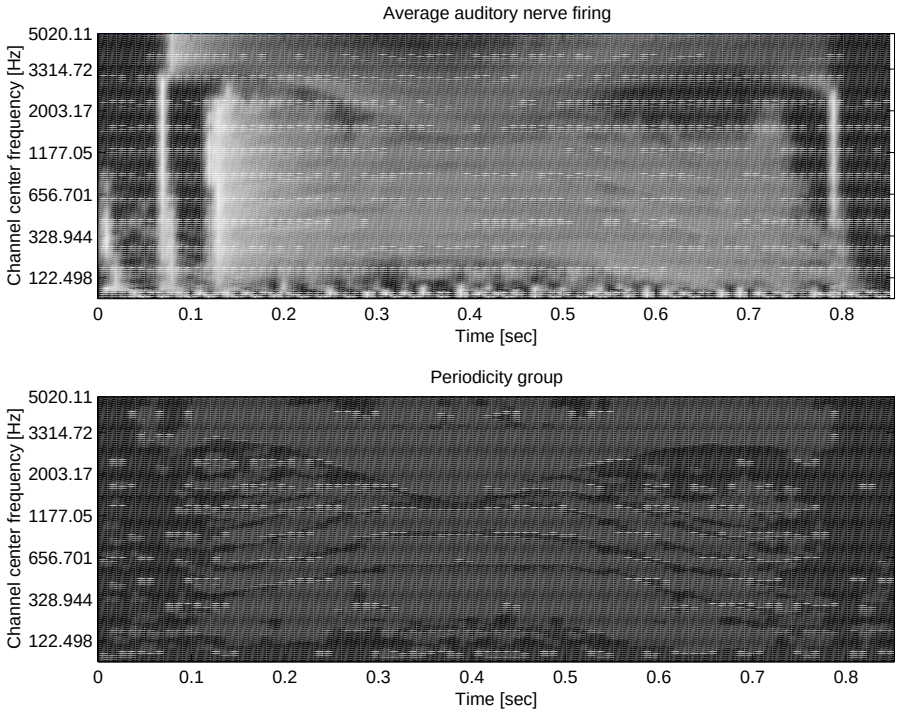

図 8: 平均発火頻度マップと Periodicity group

3.3.1 Auditory Elelement の形成

auditory element（以下 AE と記す）は periodicity group と周波数遷移マップを使って、periodicity group 上のスペクトルピークをトレースすることによって形成される。AE は、時間・周波数平面上の記号であり、たとえば個々の調波成分やフォルマントは、それぞれ 1 つの AE を形成する。

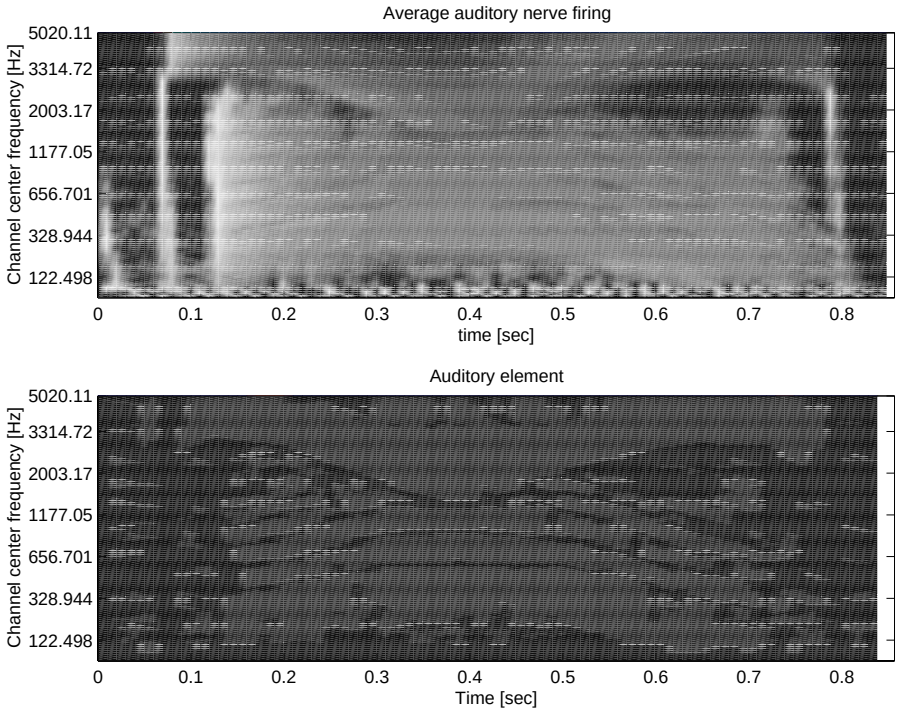

図 9: 平均発火頻度マップと Auditory element

3.3.2 Auditory Scene の探索

auditory scene の探索は、各 AE をグルーピングすることによって達成される。まず、自己相関マップ $a_n(t, f, \Delta t)$ を使って、各 AE（$f_1 \leq f \leq f_2$）毎に、時刻 t での基本周期 Δt 成分の支配確率（式 (19)）を求める。

$$P_r(t, f_1, f_2, \Delta t) = l(t, f_1, f_2, \Delta t) s_w(t, \Delta t) \tag{19}$$

ここで、

$$s_w(t, \Delta t) = \frac{w(\Delta t)}{M} \sum_{f=1}^{M} a_n(t, f, \Delta t) \tag{20}$$

$$w(\Delta t) = 1.0 - 0.9 \frac{\Delta t}{\Delta t_{max}} \tag{21}$$

$$l(t, f_1, f_2, \Delta t) = \frac{1}{f_2 - f_1 + 1} \sum_{f=f_1}^{f_2} a_n(t, f, \Delta t) \tag{22}$$

$$\tag{23}$$

次に、ダイナミックプログラミングによって、各 AE 毎に尤もらしいピッチ周期軌跡 $p_i(t)$ を求める。時刻 t における周期 $Deltat$ に対するスコア（式 (25)）は、1 時刻前のスコアと、1 時刻前の周期 $Deltat_p$ から現時刻の周期 $Deltat$ に遷移に対する遷移スコア（式 (26)）との和として定義する。

$$m(t, \Delta t) = \begin{cases} P_r(t, f_1, f_2, \Delta t) & \text{if } \frac{\partial}{\partial \Delta t} P_r(t, f_1, f_2, \Delta t) = 0 \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

$$ds(t,\Delta t) = \begin{cases} ds(t-1,\Delta t_p) + \max_{\Delta t} ts(\Delta t_p,\Delta t,t) & \text{if } t_s < t \le t_e \\ 0 & \text{if } t = t_s \end{cases} \quad (25)$$

$$ts(\Delta t_p,\Delta t,t) = m(t,\Delta t) exp\Big(-\frac{(\Delta t-\Delta t_p)^2}{2\delta t^2}\Big) \quad (26)$$

ピッチ周期軌跡が求まったら、もっとも継続時間長の長い AE を選び、新しいグループを作る。そのピッチ周期の軌跡と、その AE と時間的に重なった（$t_1 \le t \le t_2$）別の AE のピッチ周期の軌跡の類似度（式 (28)）を求める。この指標は、軌跡が厳密に一致したとき 1 となり、最小値は 0 である。次に、オンセットマップとオフセットマップ $o(t,f)$ から、式 (27) によってオンセットとオフセットの有無を調べる。式 (27) が 0 より大きくなった時点をオンセットあるいはオフセットの時刻とする。オンセットあるいはオフセットの時刻が一致した場合、ピッチ周期の軌跡の類似度にボーナスポイントとして 0.5 を加味し、合計が 0.9 を超えたとき、auditory scene から分離し、グループに加える。この一連の手続きを、すべての AE が分離されるまで行う。

$$act(t) = \sum_{f=f_1}^{f_2} \sum_{\tau=-2}^{2} o(t+\tau, f) \quad (27)$$

$$sim(p_1,p_2) = \frac{1}{t_2-t_1+1} \sum_{t=t_1}^{t_2} \exp\Big(-\frac{[p_1(t)-p_2(t)]^2}{2{\delta_p}^2}\Big) \quad (28)$$

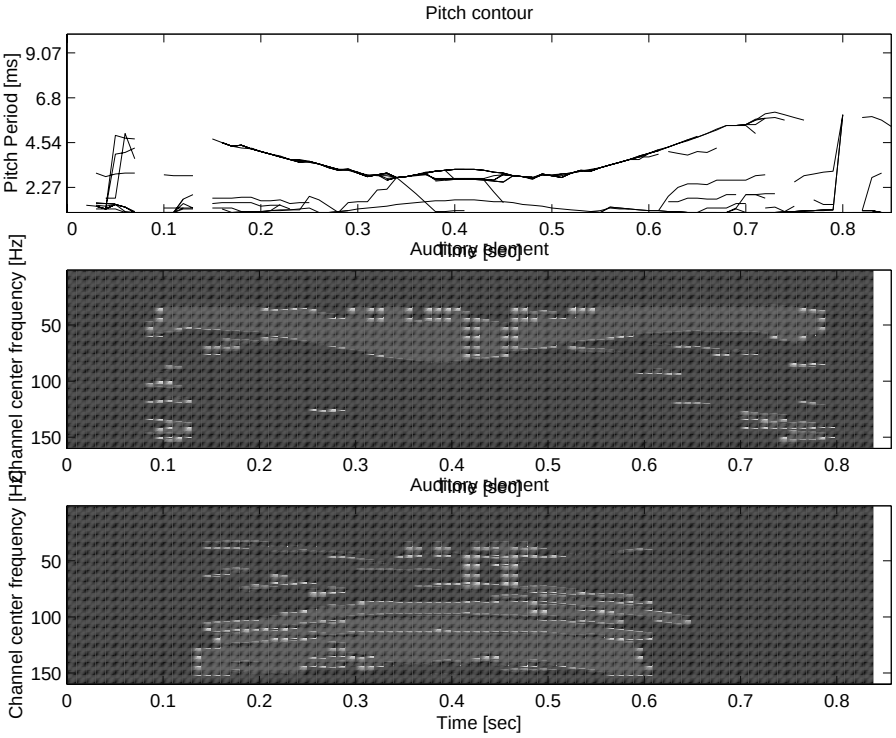

図 10: 平均発火頻度マップと F0 軌跡

4 本手法の問題点と限界

Meddis の発火モデルは、過渡特性による不要なピークが現れるなど、不安定な挙動を示すため、調波構造が乱れることなどから、auditory scene の表現には不向きである。その様子は、周波数遷移マップ（図 5）に見てとれる。また、本方法の本質的な問題点として、以下のような点が挙げられる。

- auditory element が排他的に割り当てられるので、妨害音を除去した後にスペクトル上にギャップが生じ、再生音が歪む。

- 時間的にも周波数的にも重なった auditory scene は、分離できない。
- 時間的にギャップのある auditory scene は、同一音源からの音であっても、別の auditory scene として分離される。

5 まとめ

聴覚による情景分析の考え方に基づく、混合音の音源分離手法のひとつについて紹介した。本方法は、ストリームを記述するさまざまなパラメータ（スペクトル、ピッチ、自己相関、相互相関などを時間周波数解析により求め、調波性、同期性に注目して、これらのパラメータをグルーピング手法であり、ステージ毎に閾値判定を繰り返す。当然、閾値の与え方によって、グルーピングが異なる。しかし、人の音源分離には本質的に閾値依存であり、意識的あるいは無意識的にグルーピングは変動する。その意味で、本手法は、人の音源分離の本質を捕らえていると考えられる。しかし、先に挙げたように問題も多く、さらに別の仕組みを導入する余地がある。なお、聴覚末梢系や聴覚マップのインプリメンテーションでは、Slaney[10] による Auditory Toolbox が非常に役立ったことを付記する。

参考文献

[1] A.S.Bregman,P.A.Ahad,"Auditory Scene Analysis:*The Perceptual Organization of Sound*,"The MIT Press,London,1990

[2] G.J.Brown,M.Cooke,"Computational auditory scene analysis," Computer Speech and Language,8,297-336,1994

[3] G.J.Brown,M.Cooke,"Computational auditory scene analysis: Exploiting principles of percieved continuity,"Speech Communication,13,391-399,1993

[4] E.de.Boer,H.D.Jongh,"On cochlear encoding:potentialities and limitations of the reverse correlation technique," Journal of the Acousitic Society of America,63,115-135,1978

[5] T.Irino,R.D.Patterson,"A time-domain level dependent auditory filter: The gammachirp," Journal of the Acousitic Society of America,101,412-419,1997

[6] Tatsuya.Hirahara,"Adaptive-Q filter",The Journal of the acoustical society of Japan,vol.51,7,pp.565-571,*in Japanese*

[7] B.R.Glassberg,B.C.J.Moore,"Derivation of auditory filter shapes from notched noise data," Hearing Research.,47,103-138

[8] R.Meddis,"Simulation of auditory-neural transduction:Further studies," Journal of the Acousitic Society of America 83,1056-1063,1986

[9] R.Meddis,"Implementation details of a computation model of the inner hair-cell/auditory-nerve synapse," Journal of the Acousitic Society of America, 87(4),1813-1816,1990

[10] M.Slaney,"Auditory Toolbox," Apple Technical Report #45

[11] 河原英紀,"音声コミュニケーションにおける聴覚的情景分析," 平成 6 年日本音響学会秋季研究発表会講演論文集,535-538