位相とスペクトルに着目した
聴覚の情景解析に関する基礎的研究

鵜木　祐史

1996 年 2 月 15 日

要約

キーワード：聴覚の情景解析、 2 波形分離、Gammatone filter、wavelet 分析合成系、共変調マスキング解除 (CMR)

1　はじめに

近年、Auditory Scene Analysis に基づく音源分離の研究が盛んに行なわれるようになった。音源分離問題を解くことができれば、雑音中の特定の音を聴きとるという点でカクテルパーティ効果のモデル化も可能になり、雑音に強い音声認識にも応用できる。また、聴覚の計算理論の構築に向けて新たな視点を提供してくれる。

計算機モデルの研究として、Sheffield 大学の Brown と Cooke らによる音響イベントに着目した分凝モデル、MIT の Ellis による心理音響的グルーピング規則を取り入れた分凝モデル、NTT の中谷らによる聴覚の情景解析をマルチエージェントシステムによって実現した分凝の実装例がある。また、東大の柏野らは、 2 つの周波数成分の分離知覚に関して、スペクトログラム上の複数の特徴と分離知覚の生じる割合との定量的関係をモデル化している。しかし、いずれの計算機モデルもパワー (振幅) スペクトログラム上の分離を考えているため、 2 つの信号が同じ周波数領域の成分を含むような場合、完全に分離できているとは言い難い。

本論文では、同一周波数領域において完全に分離するためには、振幅スペクトル (パワー) の他に位相も考慮しなければならないという立場に立ち、 2 波形分離問題の解法の１つとして、雑音が付加された波形から信号波形を抽出する方法を提案する。

2　モデル構成と 2 波形分離問題の定式化

本方法のモデルは、図 1 のように構成される。ある 2 つの音響信号 $f_1(t)$ と $f_2(t)$ が信号 $f(t) = f_1(t) + f_2(t)$ に合成された状況を想定する。この混合信号は、N 個の聴覚フィルタ (Gammatone filter) で構成される wavelet 分析系により周波数分解される。ここで、

$f_1(t)$ と $f_2(t)$ は、それぞれ k 番目の分析フィルタで

$$f_1(t) \quad : \quad A_k(t)\sin(\omega_k t) \tag{1}$$

$$f_2(t) \quad : \quad B_k(t)\sin(\omega_k t + \theta_k(t)) \tag{2}$$

に周波数分解されれば、フィルタ出力 $X_k(t)$ は、

$$\begin{aligned} X_k(t) &= A_k(t)\sin\omega_k t + B_k(t)\sin(\omega_k t + \theta_k(t)) \\ &= S_k(t)\sin(\omega_k t + \phi_k(t)) \end{aligned} \tag{3}$$

と表される。但し、ω_k はフィルタの中心角周波数、$\theta_k(t)$ は $f_2(t)$ のもつ $f_1(t)$ に対応した入力位相である。また、振幅包絡 $S_k(t)$ と出力位相 $\phi_k(t)$ は、それぞれ

$$S_k(t) = \sqrt{A_k^2(t) + 2A_k(t)B_k(t)cos\theta_k(t) + B_k^2(t)} \tag{4}$$

$$\phi_k(t) = \tan^{-1}\left(\frac{B_k(t)\sin(\theta_k(t))}{A_k(t) + B_k(t)\cos\theta_k(t)}\right) \tag{5}$$

となる。ここで、振幅包絡 $S_k(t)$ と中心周波数 ω_k が観測可能であることから、出力位相 $\phi_k(t)$ と入力位相 $\theta_k(t)$ がわかれば、 2 波形の振幅包絡 $A_k(t), B_k(t)$ を

$$A_k(t) = \frac{S_k(t)\sin(\theta_k(t) - \phi_k(t))}{\sin\theta_k(t)} \tag{6}$$

$$B_k(t) = \frac{S_k(t)\sin(\phi_k(t))}{\sin\theta_k(t)} \tag{7}$$

のように解析的に解くことができる。最後に、すべての分析フィルタ $(X_k(t), 1 \le k \le N)$ について、この処理を行ない、それぞれの成分を wavelet 合成系で合成することで、$f_1(t)$ と $f_2(t)$ を再構成できる。

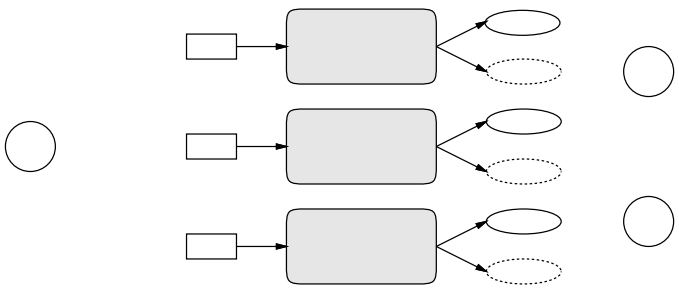

図 1: wavelet 分析合成系

尚、Gammatone filter を基底関数とした wavelet 分析合成系は、中心周波数 $f_0 = 600$[Hz]、チャネル数 $N = 129$、各分析フィルタの矩形帯域幅が重複せず (約 $\frac{1}{4}ERB$)、全周波数解析範囲 ($60 \sim 6000$[Hz]) を完全に被覆するように設計されている。

3　物理パラメータの導出方法

振幅包絡 $S_k(t)$ と出力位相 $\phi_k(t)$ は、それぞれ、複素 wavelet 変換で定義された振幅スペクトルと位相スペクトルから導出できる。また、入力位相 $\theta_k(t)$ は、(i) 漸近的変化の規則と (ii) 連続性 (近接) の規則、(iii) 1 つの音響事象に生じる変化の規則を物理的制約条件に捉え直すことで得られる。特に、規則 (i),(iii) は、Bregman が提唱した発見的規則である。

はじめに、(i) を "微小区間で $dA_k(t)/dt = 0$" という物理的制約条件として捉え直すことで、1 階線形微分方程式が得られ、この一般解は

$$\theta_k(t) = \tan^{-1}\left(\frac{S_k(t)\sin\phi_k(t)}{S_k(t)\cos\phi_k(t) + C_0}\right) \tag{8}$$

となる。但し、C_0 は未定係数である。次に、(ii) を "分離境界 $(t = T_r)$ において、3 つの物理パラメータ $Q_k(t)(= A_k(t)、B_k(t)、\theta_k(t))$ はある幅 $\Delta Q(= \Delta A、\Delta B、\Delta\theta)$ で接合されなければならない" :

$$\left|[Q_k(t)]_{R:t\to+T_t} - [Q_k(t)]_{R:t\to-T_r}\right| \le \Delta Q \tag{9}$$

という制約に捉え直すことで、未定係数 C_0 の取り得る範囲を $C_\alpha \le C_0 \le C_\beta$ に限定する。最後に、(iii) を "振幅包絡 $B_k(t)$ は、隣接する聴覚フィルタから得られる $B_{k\pm1}(t)$ に強い相関がなければならない" という制約に捉え直し、帯域雑音の振幅間の相関が最も強くなるときの未定係数 C_0 を選ぶことで、入力位相 $\theta_k(t)$ を一意に求めることができる。

$$\max_{C_\alpha \le C_0 \le C_\beta}\left(\frac{< B_k, \hat{B}_k >}{||B_k||||\hat{B}_k||}\right) \tag{10}$$

但し、$\hat{B}_k(t) = (B_{k+1}(t) + B_{k-1}(t))/2$ である。この規則 (iii) は、共変調マスキング解除 (CMR) のよい説明になっている。

2 波形分離の手順は、$(\mathbf{1})dS_k(t)/dt$ と $d\phi_k(t)/dt$ から検出可能な立ち上がりと立ち下がりによって分離区間を決定し、$(\mathbf{2})$ これを微小区間 M/f_0 に分割し、$(\mathbf{3})$ 式 (6) ~ (10) を適用して解くことである。

4　2 波形分離のシミュレーション

実験データとして、図 2 のような定常音 $f_1(t)$ と振幅変調された帯域雑音 $f_2(t)$、ランダム帯域雑音 $f_3(t)$ の 3 つの信号を用意する。但し、純音の周波数は 600[Hz]、帯域雑音は、中心周波数を 600[Hz]、帯域幅を 1[kHz] とし、純音と帯域雑音の SN 比は $-8.51[dB]$ である。また、分離に必要な各パラメータは、微小区間を $M/f_0 = 3/f_0$、$\Delta B = 50$、$\Delta\theta = \pi/20$ とし、ΔA は分離前後の $A_k(t)$ の差分値とした。ここで、CMR を想定した混合波形 $f_C(t) = f_1(t) + f_2(t)$ では、図 3 のように分離でき、再構成された信号波形の時間領域における SN 比は、$\hat{f}_1(t)$ が 15.82[dB]、$\hat{f}_2(t)$ が 11.76[dB] であった。この結果から、純音を高い精度で抽出でき、更に雑音も高い精度で分離できることがわかる。

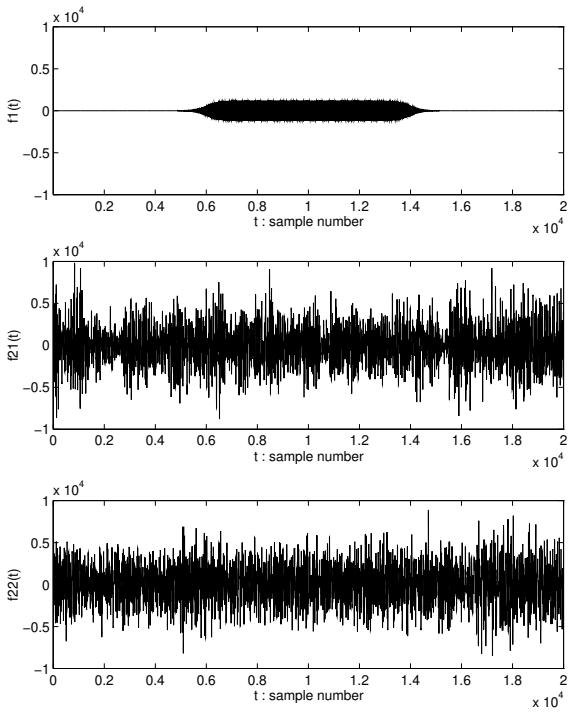

図 2: 音響信号：$f_1(t)$(上)、$f_2(t)$(中)、$f_C(t)$(下)

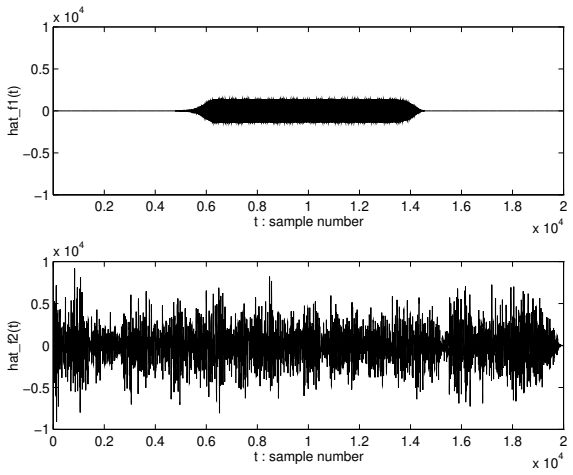

図 3: 再構成された信号：$\hat{f}_1(t)$ と $\hat{f}_2(t)$

次に、CMR の現象を説明するために、図 1 の wavelet 分析合成系を人間の聴覚特性に合わせてパラメータ設定 (フィルタの帯域幅を 1ERB、矩形帯域幅は重複可) した場合の純音の抽出結果を表 1 に示す。共変調マスキング解除を想定した混合信号 $f_C(t)$ の場合、隣接する聴覚フィルタの参照数を増加させると、抽出された純音 $\hat{f}_1(t)$ の SN 比が向上する傾向がみられた。これに対し、先の $f_2(t)$ を $f_3(t)$ に置き換えることでマスキングの状況を想定した混合信号 $f_M(t) = f_1(t) + f_3(t)$ の場合、参照数を増加させても、あまり SN 比は変わらなかった。この結果は、CMR の工学的な説明として解釈できる。

表 1: 隣接する聴覚フィルタ数–$\hat{f}_1(t)$ の SN 比の関係

隣接する聴覚フィルタ数	帯域幅 [Hz]	$f_C(t)$ SN 比 [dB]	$f_M(t)$ SN 比 [dB]
1	134	0.49	2.36
3	220	5.87	1.83
5	308	10.03	1.48

5 まとめ

雑音が付加された信号を wavelet 分析合成系で周波数分解し、振幅包絡と出力位相、入力位相の変化に着目することで、原信号を抽出する方法を提案した。本方法では、AM 帯域雑音が純音に付加された場合、純音の抽出が容易になり、ランダム帯域雑音が付加された場合、困難になることを示した。