聴覚の情景解析に基づいた二波形分離モデルの提案

鵜木　祐史　　赤木　正人

北陸先端科学技術大学院大学　情報科学研究科
〒 923-1292　石川県能美郡辰口町旭台 1-1
E-mail: unoki@jaist.ac.jp　akagi@jaist.ac.jp

本論文では、聴覚の情景解析に基づいた音源分離のモデル化の試みとして、二波形分離問題を取り上げ、雑音が付加された信号から望みの信号を分離抽出する方法を提案する。本方法では、信号の特徴として分析フィルタ群を通過した混合信号の瞬時振幅と瞬時位相を利用し、Bregman によって提唱された四つの発見的規則を制約条件として利用することで、望みの信号の瞬時振幅と瞬時位相を一意に求める。計算機シミュレーションの結果、本方法を利用することで、雑音中から実音声 (単母音と連続母音) を波形レベルにおいて高い精度で分離抽出できることが示された。また、本方法で利用する制約条件のいくつかを順番に省略した場合の分離精度を比較検討した結果、四つの発見的規則をすべて利用することの有効性が示された。

聴覚の情景解析、Bregman の発見的規則、二波形分離問題、二波形分離モデル

A model of the problem of segregating two acoustic sources based on auditory scene analysis

Masashi Unoki　and　Masato Akagi

School of Information Science,
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Nomigun, Ishikawa 923-1292 Japan
E-mail: unoki@jaist.ac.jp　akagi@jaist.ac.jp

This paper proposes a method of extracting the desired signal from a noisy signal, addressing the problem of segregating two acoustic sources as a model of acoustic source segregation based on Auditory Scene Analysis. The proposed method uses the instantaneous amplitude and phase of noisy signal components that have passed through a wavelet filterbank as features of acoustic sources. Then the model can extract the instantaneous amplitude and phase of the desired signal using the four heuristic regularities proposed by Bregman as constraints. Simulations were performed to segregate the harmonic complex tone (vowel and continuous vowel) from a noise-added harmonic complex tone and to compare the results of using all or only some constraints. The results show that the method can segregate the harmonic complex tone precisely using all the constraints related to the four regularities proposed by Bregman and that absence of some constraints reduces accuracy.

auditory scene analysis, Bregman's regularities, segregation problem, segregation model

1　まえがき

話し声や雑音など様々な音が混在する中で、望みの信号を分離抽出するという課題は、数理工学的な処理として実現するには難しい問題である。これは、混合信号から目的の音を分離抽出するという問題において、個々の音がどのように混合されたのかを表す情報が欠落していることに起因する。すなわち、この信号分離問題は、観測された信号から個々の信号を推定する不良設定の逆問題となっている。そのため、音や環境に対する何らかの制約条件がない限り、この問題を一意に解くことは難しい。

一方、聴覚は我々が経験する環境の中でいともたやすく目的の音を分離抽出できる。最近、この聴覚の優れた能力は、聴覚が能動的に外界を把握するための機能であると考えられるようになった。これは、聴覚の情景解析 (Auditory Scene Analysis: ASA) と呼ばれ、Bregman の著書 [1] により広く知られるようになった。Bregman は、音を通じて環境を把握する ASA の問題を解くために、聴覚が利用している制約条件のいくつかを、四つの発見的規則：(i) 共通の立上り/立下りに関する規則、(ii) 漸近的変化に関する規則、(iii) 調波関係に関する規則、(iv) 一つの音響事象に生じる変化に関する規則、としてまとめた [2]。

本論文は、目的音を分離抽出する問題 (音源分離問題) を ASA の問題の一つと考え、四つの発見的規則を制約条件として利用することで、不良設定の逆問題となる音源分離問題を一意に解く方法を提案するものである。この方法で音源分離を数理工学的に実現することは、音声認識システムのフロントエンドとしての応用に期待できるだけでなく、様々な聴覚心理現象のモデル化に役立てることができる。また、聴覚の計算理論の構築に向けても新たな視点を提供できるものと考えられる [3]。

工学的に情景解析 (音源分離) 問題を解こうとする研究は、計算論的な聴覚の情景解析と呼ばれ、多くの計算モデルが提案されている。これには、大きく分けて、ボトムアップ処理に基づくモデル [4, 5, 6, 7] とトップダウン処理に基づくモデル [8, 9] がある。これらのモデルのほとんどは、Bregman によって提唱された発見的規則 (i) と (iii) を利用したものであり、音響的な特徴として振幅 (もしくはパワー) スペクトルを用いている。そのため、これらのモデルでは、二つの信号が同じ周波数領域の成分を含むような場合、二つの信号を完全に分離できているとは言い難い。

著者らは、二つの信号成分が同一周波数領域に存在するとき、それらを完全に分離するためには振幅スペクトルの他に位相も考慮しなければならないと考え、基本的な音源分離問題として二波形分離問題に取り組んできた [10]。その結果、発見的規則 (ii) と (iv) を利用することで雑音中の正弦波信号を正確に分離抽出することが可能となった [10]。しかし、この方法には、(1) 二波形の瞬時位相の決定と (2) 調波複合音の分離抽出問題への拡張といった課題が残されていた。課題 (2) については、残りの規則 (i) と (iii) を利用することで、人工的な調波複合音の瞬時振幅を分離抽出することが可能となった [11] が、(1) の課題は残ったままであった。

本論文では、二波形分離問題の枠組を示し、課題 (1) の解決法を提案する。そして、雑音が付加された調波複合音 (母音) から望みの調波複合音 (母音) を分離抽出する方法を提案する。

2　二波形分離モデル

本論文では、“ある二つの独立な音源で生じた音響信号が加算された信号から、それぞれの音響信号に分離すること” を二波形分離問題と定義する。この問題は以下のように定式化される [10]。

2.1　二波形分離問題の定式化

はじめに、ある二つの音響信号 $f_1(t)$ と $f_2(t)$ が $f(t) = f_1(t) + f_2(t)$ に加算され、混合信号 $f(t)$ のみを受音できるものとする。ここで、$f_1(t)$ を望みの信号、$f_2(t)$ を雑音あるいはそれ以外の音とする。これは、図 1に示す K 個の分析フィルタ群により周波数分解される。ここで、k 番目の分析フィルタを通過した $f_1(t)$ と $f_2(t)$ の周波数成分を、それぞれ

$$X_{1,k}(t) = A_k(t)\exp(j\omega_k t + j\theta_{1k}(t)) \quad (1)$$
$$X_{2,k}(t) = B_k(t)\exp(j\omega_k t + j\theta_{2k}(t)) \quad (2)$$

と仮定すれば、$f(t)$ の通過成分 $X_k(t)$ は、

$$X_k(t) = X_{1,k}(t) + X_{2,k}(t) \quad (3)$$
$$= S_k(t)\exp(j\omega_k t + j\phi_k(t)) \quad (4)$$

と表される。但し、ω_k は分析フィルタの中心角周波数、$A_k(t)$、$B_k(t)$、$S_k(t)$ は瞬時振幅、$\phi_k(t)$ は瞬時出力位相、$\theta_{1k}(t)$ と $\theta_{2k}(t)$ は瞬時入力位相である。また、$S_k(t)$ と $\phi_k(t)$ は、それぞれ、式 (1) ~ (4) から

$$S_k(t) = \sqrt{A_k^2(t) + 2A_k(t)B_k(t)\cos\theta_k(t) + B_k^2(t)} \quad (5)$$

$$\phi_k(t) = \arctan\left(\frac{A_k(t)\sin\theta_{1k}(t) + B_k(t)\sin\theta_{2k}(t)}{A_k(t)\cos\theta_{1k}(t) + B_k(t)\cos\theta_{2k}(t)}\right) \quad (6)$$

で求められるため、$A_k(t)$ と $B_k(t)$ は、それぞれ

$$A_k(t) = \frac{S_k(t)\sin(\theta_{2k}(t)-\phi_k(t))}{\sin\theta_k(t)} \quad (7)$$

$$B_k(t) = \frac{S_k(t)\sin(\phi_k(t)-\theta_{1k}(t))}{\sin\theta_k(t)} \quad (8)$$

として解くことができる。但し、$\theta_k(t)=\theta_{2k}(t)-\theta_{1k}(t)$ であり、$\theta_k(t)\neq n\pi, n\in\mathbf{Z}$ とする。同様に上式を整理すると、$\theta_{1k}(t)$ と $\theta_{2k}(t)$ は、それぞれ

$$\theta_{1k}(t) = -\arctan\left(\frac{Y_k(t)\cos\phi_k(t)-\sin\phi_k(t)}{Y_k(t)\sin\phi_k(t)+\cos\phi_k(t)}\right) + \arcsin\left(\frac{A_k(t)Y_k(t)}{S_k(t)\sqrt{Y_k(t)^2+1}}\right) \quad (9)$$

$$\theta_{2k}(t) = -\arctan\left(\frac{Y_k(t)\cos\phi_k(t)+\sin\phi_k(t)}{Y_k(t)\sin\phi_k(t)-\cos\phi_k(t)}\right) + \arcsin\left(-\frac{B_k(t)Y_k(t)}{S_k(t)\sqrt{Y_k(t)^2+1}}\right) \quad (10)$$

として解くことができる。但し、

$$Y_k(t)=\sqrt{(2A_k(t)B_k(t))^2-Z_k(t)^2}/Z_k(t) \quad (11)$$

$$Z_k(t)=S_k(t)^2-A_k(t)^2-B_k(t)^2 \quad (12)$$

である。しかし、上記の定式化において、観測された混合信号の $S_k(t)$ と $\phi_k(t)$ から、四つのパラメータ ($A_k(t)$、$B_k(t)$、$\theta_{1k}(t)$、$\theta_{2k}(t)$) を同時に、かつ一意に求めることはできない。これは 二波形分離問題が不良設定の逆問題であることに起因している。本定式化は、sinusoidal model [12] の定式化に等しいため、例えばピークピッキングやハーモニックトラッキングの方法 [12] などで各パラメータを推定することができる。しかし、これらの方法では、二波形が同一周波数範囲に存在するとき、二波形の各パラメータを正確に推定することはできない。

そこで、本論文では、Bregman によって提唱された四つの発見的規則を制約条件として用いて、分離抽出したい信号を拘束することで、$S_k(t)$ と $\phi_k(t)$ から四つのパラメータを一意に求めることを考える。

2.2 モデルで利用する制約条件

本論文では、$f_1(t)$ を調波複合音と仮定し、雑音 $f_2(t)$ 中に $f_1(t)$ が加算される状態から $f_1(t)$ を分離抽出する問題を扱う。また、この調波複合音は、基本周波数 $F_0(t)$ を整数倍した高調波成分をもつものとする。

先の論文 [11] では、四つの発見的規則を数理工学的な制約条件として利用し、分離抽出したい信号の瞬時振幅と基本周波数の時間変化を拘束することで二波形分離問題の最適解を求めた。この方法では、

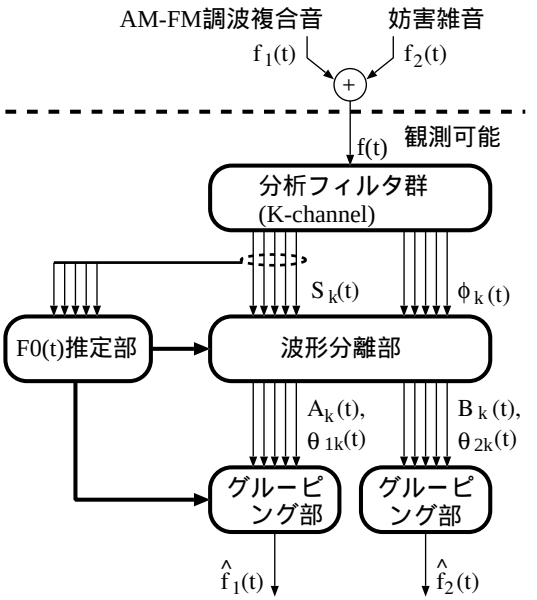

図 1: 二波形分離モデル

調波複合音の瞬時振幅を正確に分離抽出することができたが、瞬時位相については十分正確に分離抽出することができなかった。

本論文では、分離抽出したい信号の瞬時振幅と基本周波数の時間変化の他に、瞬時位相の時間変化にも着目して、望みの信号を分離抽出する。そこで、先の論文 [10, 11] で定義した制約条件を再度整理し、次のように再定義する。

制約条件 1 (漸近的変化 (多項式近似)) ある区間における瞬時振幅 $A_k(t)$、瞬時入力位相 $\theta_{1k}(t)$、基本周波数 $F_0(t)$ それぞれの導関数が、

$$dA_k(t)/dt = C_{k,R}(t) \quad (13)$$

$$d\theta_{1k}(t)/dt = D_{k,R}(t) \quad (14)$$

$$dF_0(t)/dt = E_{0,R}(t) \quad (15)$$

で表されるものとする。但し、$C_{k,R}(t)$, $D_{k,R}(t)$, $E_{0,R}(t)$ は、区分的に R 回微分可能な R 次多項式である。このとき、$A_k(t)$, $\theta_{1k}(t)$, $F_0(t)$ は、それぞれ、$A_k(t)=\int C_{k,R}(t)dt+C_{k,0}$, $\theta_{1k}(t)=\int D_{k,R}(t)dt+D_{k,0}$, $F_0(t)=\int E_{0,R}(t)dt+E_{0,0}$ と表される。 □

制約条件 2 (調波関係) 基本周波数を $F_0(t)$、高調波の次数を N_{F_0} とする。このとき、調波関係にある信号成分は

$$n\times F_0(t), \qquad n=1,2,\cdots,N_{F_0} \quad (16)$$

の関係を満たさなければならない。 □

制約条件 3 (立上り・立下りの同期) 基本波成分の立上り時刻を T_{S}、立下り時刻を T_{E} とする。この

とき、同じ音源で生じた信号成分であれば、k 番目の高調波成分の立上り $T_{k,\mathrm{on}}$ と立下り $T_{k,\mathrm{off}}$ は基本波の立上りと立下りに一致しなければならない。すなわち、それぞれの一致の誤差は

$$|T_{\mathrm{S}} - T_{k,\mathrm{on}}| \leq \Delta T_{\mathrm{S}} \quad (17)$$
$$|T_{\mathrm{E}} - T_{k,\mathrm{off}}| \leq \Delta T_{\mathrm{E}} \quad (18)$$

を満たさなければならない。 □

制約条件 **4** (漸近的変化 (なめらかさ)) 閉区間 $[t_a, t_b]$ における $A_k(t)$ と $\theta_{1k}(t)$ に対し、定積分

$$\sigma_A = \int_{t_a}^{t_b} [A_k^{(R+1)}(t)]^2 dt \quad (19)$$
$$\sigma_\theta = \int_{t_a}^{t_b} [\theta_{1k}^{(R+1)}(t)]^2 dt \quad (20)$$

が最小になるとき、$A_k(t)$ および $\theta_{1k}(t)$ を最もなめらかであるとする。但し、$A_k(t)$ と $\theta_{1k}(t)$ は、それぞれ、式 *(13)* の $C_{k,R}(t)$ と式 *(14)* の $D_{k,R}(t)$ を用いて決定された瞬時振幅と瞬時位相である。また、$A_k^{(R+1)}(t)$ と $\theta_{1k}^{(R+1)}(t)$ は、それぞれ $A_k(t)$ と $\theta_{1k}(t)$ の $(R+1)$ 次導関数である。 □

制約条件 **5** (振幅包絡 $A_k(t)$ 間の相関) 振幅包絡 $A_k(t)$ は隣接する ℓ 番目の分析フィルタにおける振幅包絡 $A_\ell(t)$ に強い相関がなければならない :

$$\frac{A_k(t)}{\|A_k(t)\|} \approx \frac{A_\ell(t)}{\|A_\ell(t)\|}, \qquad k \neq \ell. \quad (21)$$

但し、$\|\cdot\|$ はノルム記号である。 □

ここで、制約条件式 (13) を式 (7) に適用することで、一階線形微分方程式が得られる。これを解くことで入力位相差 $\theta_k(t)$ の一般解を

$$\theta_k(t) = \arctan\left(\frac{S_k(t)\sin(\phi_k(t) - \theta_{1k}(t))}{S_k(t)\cos(\phi_k(t) - \theta_{1k}(t)) - C_k(t)}\right) \quad (22)$$

で得ることができる。但し、$C_k(t) = \int C_{k,R}(t)dt + C_{k,0}$ である。従って、$C_{k,R}(t)$ が決まれば、式 (22) から $\theta_k(t)$ を決定でき、$D_{k,R}(t)$ が決まれば、$\theta_{1k}(t)$ と $\theta_{2k}(t)$ を決定できる。

2.3 モデルの信号処理の概要

本論文で提案する二波形分離問題の解法は、図 1 に示すモデルで実現される。このモデルは、(a) 分析フィルタ群、(b) 基本周波数の推定部、(c) 波形分離部、(d) グルーピング部の 4 ブロックで構成される。

はじめに、混合信号 $f(t)$ のみが観測され、分析フィルタ群により、$S_k(t)$ と $\phi_k(t)$ に分解される。次に、$S_k(t)$ から基本周波数 $F_0(t)$ を求め、二波形分離の対象となる時間–周波数領域を決定する。調波成分の存在する周波数領域については、$F_0(t)$ と発見的規則 (iii) の調波関係 (制約条件 2) を用いて決定する。調波成分の存在する時間領域については、発見的規則 (i) の、各高調波成分の立上りと立下りの同期 (制約条件 3) を用いて決定する。

次に、波形分離部では、上記で決定された時間–周波数領域において $S_k(t)$ と $\phi_k(t)$ から $A_k(t)$, $B_k(t)$, $\theta_{1k}(t)$, $\theta_{2k}(t)$ を求める。これは、$A_k(t)$ と $\theta_{1k}(t)$ を発見的規則 (ii) の漸近的変化 (ゆっくりと) を用いて最適化問題として解く。但し、最適解の候補が多過ぎるため、発見的規則 (ii) の漸近的変化 (なめらかさ) を加え、解の探索範囲を狭め、発見的規則 (iv) の変動の一致 (相関) を手がかりとして最適解の絞り込みを行う。

最後に、グルーピング部では、$A_k(t)$ と $\theta_{1k}(t)$、および $B_k(t)$ と $\theta_{2k}(t)$ がグルーピングされ、分析フィルタ群によるフィルタリング処理とは逆の操作を行うことで、それぞれ $\hat{f}_1(t)$ と $\hat{f}_2(t)$ に再構成される。

3 アルゴリズムの実装

3.1 分析合成系の実装

本論文では、(1) 聴覚特性を考慮できること、(2) 複素スペクトルを扱えて、不連続点の検出が容易であることを考慮して、gammatone filter をアナライジング wavelet とした wavelet 分析合成系 (定 Q フィルタバンク) [10] を利用する。この分析合成系は、サンプリング周波数 f_s を 20 kHz、通過帯域幅を 60 ~ 6000 Hz、チャンネル数 K を 128 として設計された [10]。ここで、式 (5) の $S_k(t)$ と式 (6) の $\phi_k(t)$ は、それぞれ、$f(t)$ の wavelet 変換の振幅項と位相項から得ることができる [10]。

本論文では、基本周波数 $F_0(t)$ の推定方法として、分析フィルタ群の出力 ($S_k(t)$) 上における Comb filtering を採用する [11]。ここで、推定された $F_0(t)$ に対し、漸近的変化の制約条件式 (15) を考える。実装された分析フィルタ群では、分析フィルタ群の各中心周波数は離散値を取るため、$F_0(t)$ は階段状に変化する。そこで、$F_0(t)$ が変化しない区間において、式 (15) を $E_{0,R}(t) = 0$ と解釈する。ここで、階段状となる $F_0(t)$ の不連続点が H 個あるとき、その点の時刻を $T_1, T_2, \cdots, T_{H-1}, T_H$ とする。

Step. 1 Kalman filter を用いて、式 (13) の $C_{k,0}(t)$ と式 (14) の $D_{k,0}(t)$ を推定する。

Step. 2 推定誤差内 $\hat{D}_{k,0}(t) - Q_k(t) \leq D_{k,1}(t) \leq \hat{D}_{k,0}(t) + Q_k(t)$ から、Spline 補間された $D_{k,1}(t)$ の候補を求める。

Step. 3 推定誤差内 $\hat{C}_{k,0}(t) - P_k(t) \leq C_{k,1}(t) \leq \hat{C}_{k,0}(t) + P_k(t)$ から、Spline 補間された $C_{k,1}(t)$ の候補を求める。

Step. 4 式 (30) の相関値最大を尺度に、$\hat{C}_{k,1}(t)$ を求める。

Step. 5 Step.3 ~ 4 を繰り返し、式 (32) の相関値最大を尺度に、$\hat{D}_{k,1}(t)$ を決定する。

Step. 6 $\hat{C}_{k,1}(t)$ から $\theta_k(t)$ を、$\hat{D}_{k,1}(t)$ から $\theta_{1k}(t)$ を決定する。これより、$\theta_{2k}(t) = \theta_k(t) + \theta_{1k}(t)$ を決定する。

Step. 7 式 (7) と式 (8) から $A_k(t)$ と $B_k(t)$ を決定する。

図 2: パラメータの決定手順

3.2 グルーピング部の実装

各区間で一定な $F_0(t)$ に対し、制約条件 2 の調波関係と制約条件 3 の立上り・立下りの同期を実装する。

はじめに、式 (16) の制約条件 2 から、調波関係にある信号成分が存在する分析フィルタのチャネル番号を

$$\ell = \frac{K}{2} - \left\lceil \frac{\log(n \cdot F_0(t)/f_0)}{\log \alpha} \right\rceil \qquad (23)$$

で決定する。但し、$n = 1, 2, \cdots, N_{F_0}$, α は wavelet 変換のスケールパラメータであり、$\lceil \cdot \rceil$ は、正の無限大方向へ最も近い整数値への丸め記号である。また、K は偶数であり、$f_0 = 600$ Hz である。

次に、式 (17) と式 (18) の制約条件 3 から、基本波の立上りと立下りをそれぞれ、$T_{\rm S} = T_1$、$T_{\rm E} = T_H$ とする。また、一致範囲をそれぞれ $\Delta T_{\rm S} = 50$ ms と $\Delta T_{\rm E} = 100$ ms とする。$\Delta T_{\rm S}$ については、立上りの同期に関する聴取実験の結果を参考にしたものである [13]。尚、k 番目の高調波成分の立上り $T_{k,\rm on}$ と立下り $T_{k,\rm off}$ は、$S_k(t)$ と $\phi_k(t)$ の時間微分の極大点および極小点を利用して決定される [10]。

3.3 波形分離部の実装

波形分離部では、二波形分離の対象になる分析フィルタ出力において、$S_k(t)$ と $\phi_k(t)$ から、$A_k(t)$, $B_k(t)$, $\theta_{1k}(t)$, $\theta_{2k}(t)$ を決定する。本論文では、$C_{k,R}(t)$ と $D_{k,R}(t)$ の係数推定の計算量を抑えるために、制約条件 1 を $dA_k(t)/dt = C_{k,1}(t)$、$d\theta_{1k}(t)/dt = D_{k,1}(t)$ と仮定し、図 2に示す手順でこれらの係数を求める。

表 1: Kalman filtering による推定

記号	$C_{k,0}(t)$	$D_{k,0}(t)$
観測信号 $\mathbf{y}_m$	$X_k(t_m)$	$\exp(j\phi_k(t_m))$
状態変数 $\mathbf{x}_m$	$C_k(t_m)$	$\exp(jD_k(t_m))$
観測雑音 $\mathbf{v}_m$	$X_{2,k}(t_m)$	$X_{2,k}(t_m)/S_k(t_m)$
システム雑音 $\mathbf{w}_m$	w_m	w_m
状態遷移行列 $\mathbf{F}_m$	$\Delta C_k(t_m)$	$\Delta D_k(t_m)$
観測行列 $\mathbf{H}_m$	$\exp(j\omega_k t_m)$	$\hat{C}_{k,0}(t_m)/S_k(t_m)$
駆動行列 $\mathbf{G}_m$	1	1

上記の仮定の場合、$A_k(t)$ と $\theta_{1k}(t)$ は 2 次の多項式で表現できる範囲内で時間変動を許されたことになる。Step. 1 ~ 5 の処理については、次節で詳細を説明する。

3.3.1 Kalman filter を用いた推定範囲の決定

はじめに、Kalman filter を用いて $C_{k,0}(t)$ と $D_{k,0}(t)$ を推定する。ここで、$C_k(t) = \int C_{k,0}(t)dt$、$D_k(t) = \int D_{k,0}(t)dt$ とする。推定区間は基本周波数 $F_0(t)$ が一定となる一区間 $[T_{h-1}, T_h]$ である。この区間を、離散時刻 $t_m = T_{h-1} + m/f_s, m = 0, 1, \cdots, (T_h - T_{h-1})f_s$ に分割し、時刻 t_m の係数 $C_k(t_m)$ と $D_k(t_m)$ の時間変化を

$$C_k(t_{m+1}) = C_k(t_m)\Delta C_k(t_m) + w_m \qquad (24)$$

$$\Delta C_k(t_m) = 1 + \frac{C_k(t_m) - C_k(t_{m-1})}{C_k(t_m)} \qquad (25)$$

$$D_k(t_{m+1}) = D_k(t_m)\Delta D_k(t_m) + w_m \qquad (26)$$

$$\Delta D_k(t_m) = 1 + \frac{D_k(t_m) - D_k(t_{m-1})}{D_k(t_m)} \qquad (27)$$

とする。但し、$t_0 = T_{h-1}$、$t_M = T_h$ である。ここで、w_m は、平均 0 で分散 σ_w^2 の白色雑音である。

次に、式 (24) と式 (3)、式 (26) と式 (3) を $S_k(t)$ で正規化した式を、Kalman filtering 問題 [14]：

$$\mathbf{x}_{m+1} = \mathbf{F}_m\mathbf{x}_m + \mathbf{G}_m\mathbf{w}_m \quad (\text{状態方程式}) \qquad (28)$$

$$\mathbf{y}_m = \mathbf{H}_m\mathbf{x}_m + \mathbf{v}_m \quad (\text{観測方程式}) \qquad (29)$$

に対応させる。このとき、上式の各変数は、表 1 のように対応づけられる。次に、式 (28) と式 (29) に、Kalman filtering のアルゴリズム [14] を逐次適用すると、最小分散推定量 $\hat{\mathbf{x}}(t_m) = \hat{\mathbf{x}}_{m|m}$ と誤差の共分散行列 $\hat{\mathbf{e}}(t_m) = \hat{\mathbf{\Sigma}}_{m|m}$ を得る。ここで、推定値と推定誤差をそれぞれ、$\hat{C}_{k,0}(t) = |d\hat{\mathbf{x}}(t)/dt|$ と $P_k(t) = |d\hat{\mathbf{e}}(t)/dt|$、$\hat{D}_{k,0}(t) = \arg(d\hat{\mathbf{x}}(t)/dt)$ と $Q_k(t) = \arg(d\hat{\mathbf{e}}(t)/dt)$ で決定する。

3.3.2 Spline 補間を用いた候補選定

制約条件 4 におけるなめらかさの制約条件式 (19) を満たす $A_k(t)$ と制約条件式 (20) を満たす $\theta_{1k}(t)$

を求めるために、$C_{k,1}(t)$ と $D_{k,1}(t)$ の候補を選定する。ここで、ある $C_{k,R}(t)$ によって得られた瞬時振幅の $(R+1)$ 次導関数を $A_k^{(R+1)}(t)$、ある $D_{k,R}(t)$ によって得られた瞬時位相の $(R+1)$ 次導関数を $\theta_{1k}^{(R+1)}(t)$ とし、$\tau_1, \tau_2, \cdots, \tau_i$ は開区間 (t_a, t_b) に含まれ、$t_a < \tau_1 < \cdots < \tau_i < t_b$ とする。ここで、式 (19) を満たす $C_{k,R}(t),\ R=1$ と, 式 (20) を満たす $D_{k,R}(t),\ R=1$ を推定することは、閉区間 $[t_a, t_b]$ において、$A_k(\tau_i) = A_{k,i},\ \theta_{1k}(\tau_i) = \theta_{1k,i},\ i = 1, 2, \cdots, I$ となる I 個の点を通る最もなめらかな補間関数 $A_k(t),\ \theta_{1k}(t)$ を求めることに等しい [10]。この制約での最良補間関数は、$(2R+1)$ 次 Spline 関数であり、唯一存在する [15]。そこで、推定誤差範囲内で Spline 補間された $C_{k,1}(t)$ と $D_{k,1}(t)$ の各候補を求め、その候補から正しい解を一つ求めることで、最もなめらかな瞬時振幅 $A_k(t)$ と瞬時位相 $\theta_{1k}(t)$ の真の解を一意に求める。

本論文では、$R=1$ から、3 次 Spline 関数を用いて補間した。また、補間範囲は、$t_a = T_{h-1}$、$t_b = T_h$ であり、補間間隔を $\Delta\tau = 15 \times (2\pi/\omega_k)/f_s$ とした。従って、補間点数 I は、$I = \lceil (t_b - t_a)/\Delta\tau \rceil$ である。

3.3.3 相関を手がかりにしたパラメータの決定

制約条件 5 を用いて、Spline 補間された $C_{k,1}(t)$ の候補を一つの最適解に絞り込む。これは、振幅包絡 $A_k(t)$ 間の相関が、推定誤差内で最大となるときの $C_{k,1}(t)$ を選択することで実現される [11]。

$$\hat{C}_{k,1} = \underset{\hat{C}_{k,0}-P_k \le C_{k,1} \le \hat{C}_{k,0}+P_k}{\arg\max} \frac{< \hat{A}_k, \hat{\hat{A}}_k >}{||\hat{A}_k|| ||\hat{\hat{A}}_k||} \quad (30)$$

但し、$< \cdot >$ は内積記号である。ここで、$\hat{A}_k(t)$ は、ある $D_{k,0}(t)$ と Spline 補間された $C_{k,1}(t)$ により得られた振幅包絡であり、$\hat{\hat{A}}_k(t)$ は、

$$\hat{\hat{A}}_k(t) = \frac{1}{N_{F_0}} \sum_{\ell \in \mathbf{L}, \ell \neq k} \frac{\hat{A}_\ell(t)}{\|\hat{A}_\ell(t)\|} \quad (31)$$

である。但し、$\mathbf{L}$ は式 (23) を満たす ℓ の集合である。

次に、Spline 補間された $D_{k,1}(t)$ の候補を一つに絞り込む。これは、上記の手順と同様に、振幅包絡 $A_k(t)$ 間の相関を手がかりに、

$$\hat{D}_{k,1} = \underset{\hat{D}_{k,0}-Q_k \le D_{k,1} \le \hat{D}_{k,0}+Q_k}{\arg\max} \frac{< \hat{A}_k, \hat{\hat{A}}_k >}{||\hat{A}_k|| ||\hat{\hat{A}}_k||} \quad (32)$$

で、$D_{k,1}(t)$ の最適解を決定する。但し、$\hat{A}_k(t)$ は $\hat{C}_{k,1}(t)$ と Spline 補間された $D_{k,1}(t)$ により決定された振幅包絡であり、$\hat{\hat{A}}_k(t)$ は式 (31) である。

表 2: 実験データ.

Sim. No.	$f_1(t)$	$f_2(t)$
1	/a/, /i/, /u/, /e/, /o/ (mau, mht, fkn, fsu)	ピンク帯域雑音
2	/aoi/ (mau, mht, fkn, fsu)	ピンク帯域雑音

4 二波形分離のシミュレーション

本モデルが、雑音下における実音声の分離抽出において、どの程度正確に混合信号 $f(t)$ から望みの信号 $f_1(t)$ を分離抽出できるかを評価するために、(1) 雑音下の単母音の分離抽出、(2) 雑音下の連続母音の分離抽出の評価実験を行う。特に、(1) では本モデルが方略どおりに調波複合音を分離抽出できることを、(2) では基本周波数の時間的変動や調音結合の有無にかかわらず本モデルが正確に目的音を分離抽出できることを示すことが狙いである。また、評価に利用する実験データとして、ATR 音声データベースデータセット [16] にある男性 2 名 (mau, mht) と女性 2 名 (fkn, fsu) の単母音 (/a/, /i/, /u/, /e/, /o/) と連続母音 (/aoi/) を利用する。また、雑音については、帯域幅 6 kHz で帯域制限されたピンク雑音を利用する。

本論文では、分離精度の評価尺度として、$f_1(t)$ を信号、$f_1(t)$ と $\hat{f}_1(t)$ の差を雑音とみなした時間領域における SNR：

$$10 \log_{10} \frac{\int_0^T f_1(t)^2 dt}{\int_0^T (f_1(t) - \hat{f}_1(t))^2 dt} \quad \text{(dB)} \quad (33)$$

を利用する。以後、本論文では上記の SNR を分離精度 (segregation accuracy) と呼ぶ。この評価尺度を用いることで、二波形の瞬時振幅だけでなく瞬時位相も正確に分離でき、かつ正確に波形レベルに復元できることを示すことができる。

次に、本モデルで利用した制約条件の有効性を考察するために、三つの条件 [11]：

Condition 1 Comb filter による調波成分抽出 + Kalman filter で求めた $C_{k,0}(t)$ と $D_{k,0}(t)$ の利用

Condition 2 Comb filter による調波成分抽出

Condition 3 処理なし (分析合成系による全域通過)

の比較も行う。ここで、Condition 1 は、制約条件 4 のなめらかさと制約条件 5 の振幅包絡間の相関を省略した場合、Condtion 2 は更に制約条件 1 の漸近的変化を省略した場合、Condition 3 は、すべての制約条件を省略したものである。

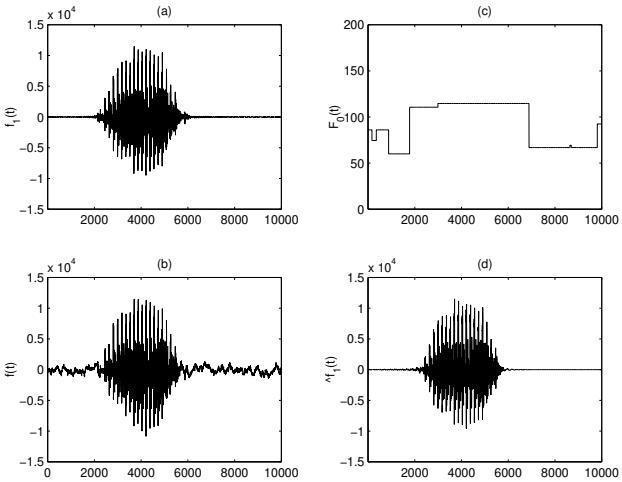

図 3: 評価実験 1 の分離例：(a) 原信号 $f_1(t)$ (mau /a/)、(b) 混合信号 $f(t)$、(c) 基本周波数 $F_0(t)$、(d) 分離抽出された信号 $\hat{f}_1(t)$

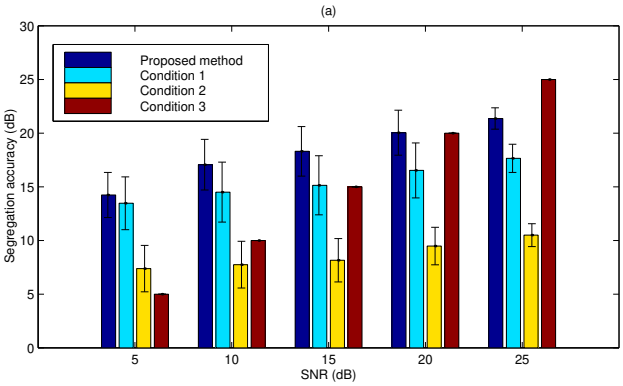

図 4: 評価実験 1 の分離精度の比較

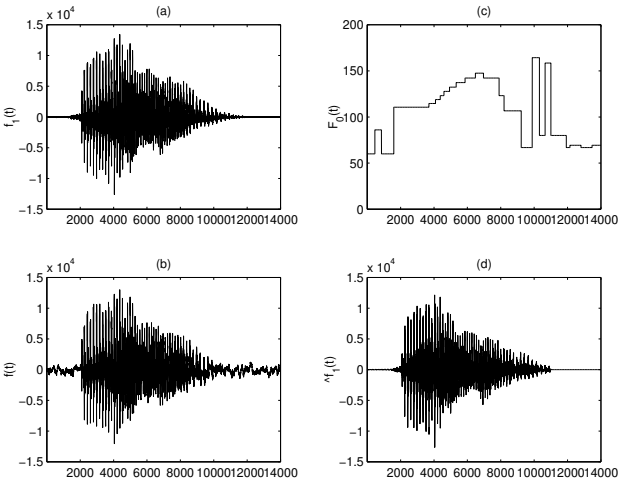

図 5: 評価実験 2 の分離例：(a) 原信号 $f_1(t)$ (mau /aoi/)、(b) 混合信号 $f(t)$、(c) 基本周波数 $F_0(t)$、(d) 分離抽出された信号 $\hat{f}_1(t)$

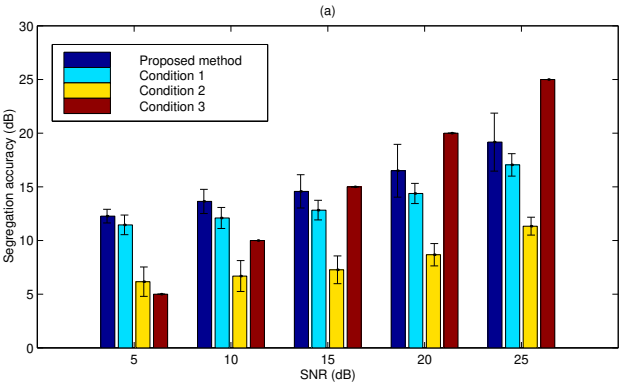

図 6: 評価実験 2 の分離精度の比較

4.1 評価実験 1 ：雑音下の単母音の分離抽出

評価実験 1 では、表 2の 1. に示す $f_1(t)$ と $f_2(t)$ の SNR を 5 dB から 25 dB まで 5 dB 刻みに変化させた、合計 100 個 (5 SNR × 4 話者× 5 母音) の混合信号 $f(t)$ を利用する。

例えば、図 3(a) に示すような $f_1(t)$ (話者 mau の母音/a/) に、SNR が 15 dB のピンク帯域雑音を付加したとき、図 3(b) に示すような $f(t)$ となる。本モデルは、図 3(c) に示すように $f(t)$ から $f_1(t)$ の基本周波数を推定し、図 3(d) に示すように、$f(t)$ から $\hat{f}_1(t)$ を 25.7 dB の精度で分離抽出できる。

次に、本モデルと三つの条件 (Condition 1、2、3) の比較を行ったところ、図 4の結果を得た。図中の棒グラフは分離精度の平均 (話者と母音の数で平均をとったもの) を、縦棒は標準偏差を示す。この図から、本モデルを利用した場合の分離精度が他の三つの条件よりも良好であることがわかる。本モデルと Condition 1 の比較では、なめらかさ (制約条件 4) の制約を利用したことによる分離精度の向上を確認できる。本モデルと Condition 2 の比較では、同一周波数領域に二波形の成分が存在する際、位相情報を利用したことによる分離精度の向上を確認できる。本モデルと Condition 3 の比較では、本モデルの分離精度の向上 (雑音除去能力) を求めることができる。この結果、$f(t)$ の SNR が 5 dB (最悪 SNR) のとき、$\hat{f}_1(t)$ の分離精度が 9.2 dB 改善されたことがわかる。

4.2 評価実験 2 :雑音下の連続母音の分離抽出

評価実験 2 では、表 2の 2. に示す $f_1(t)$ と $f_2(t)$ の SNR を 5 dB から 25 dB まで 5 dB 刻みに変化させた、合計 20 個 (5 SNR × 4 話者× 1 連続母音) の混合信号 $f(t)$ を利用する。

例えば、図 5(a) に示すような $f_1(t)$ (話者 mau の母音/aoi/) に、SNR が 15 dB のピンク帯域雑音を付加したとき、図 5(b) に示すような $f(t)$ となる。本モデルは、図 5(c) に示すように $f(t)$ から $f_1(t)$ の基本周波数を推定し、図 5(d) に示すように、$f(t)$ から $\hat{f}_1(t)$ を 17.2 dB の精度で分離抽出できる。

次に、本モデルと三つの条件 (Condition 1、2、3) の比較を行ったところ、図 6の結果を得た。図中の平均と標準偏差は図 4と同じ方法で計算したもの

である。この図から、本モデルを利用した場合の分離精度が他の三つの条件よりも良好であることがわかる。この結果、$f(t)$ の SNR が 5 dB (最悪 SNR) のとき、$\hat{f}_1(t)$ の分離精度が 7.3 dB 改善されたことがわかる。

4.3 考察

図 4と図 6の結果から、本モデル (制約条件すべて) を利用することの有効性が示された。しかし、SNR が最良時 (25 dB 以上) のとき、混合信号を原信号と見なした場合 (Condition 3 に対応) の分離精度と大差ないか、もしくは若干低下する結果となった。これは、二波形分離モデルにおける瞬時振幅、瞬時位相の時間変動を R 次の区分多項式で近似する際、計算量の削減のために $R = 1$ としたことに起因する。従って、多項式近似の次数を高くすることでこの問題点は改善されるものと考えられる。

5 おわりに

本論文では、聴覚の情景解析に基づいた音源分離問題のモデル化の試みとして、雑音が付加された調波複合音から望みの調波複合音を分離抽出する方法を提案した。この方法は、信号と雑音が混合された状態から、望みの信号を分離抽出する不良設定の逆問題を解くものである。本方法では、信号の特徴として、分析フィルタ群で分解された混合信号の瞬時振幅と瞬時位相を利用した。また、二波形分離問題の一意な解を求めるために、制約条件として Bregman によって提唱された四つの発見的規則を利用した。この方法は、先に提案した方法 [10, 11] をベースとしており、瞬時位相に対する制約を考慮することで、これまでに課題となっていた波形レベルでの分離精度の比較が可能となった。

本方法の有効性を示すために、(1) 雑音下の単母音の分離抽出、(2) 雑音下の連続母音の分離抽出、といった二つの評価実験を行った。この結果、制約条件の考察を行ったところ、二つのシミュレーションの結果すべてにおいて、四つの発見的規則を制約条件として利用することの有効性が確認された。また、同一周波数領域に二波形の成分が存在するとき、振幅の連続性だけでなく位相の連続性も考慮することで、二波形の分離精度を向上できることが確認された。

今後は、実音声を対象とした際の区分的多項式の近似次数 (R) の検証と、実環境や子音も含めた実音声の二波形分離問題への本モデルの適用が課題である。

謝辞

本研究の一部は、日本学術振興会特別研究員研究奨励金、科学研究費補助金 (07308926、10680374) および科学技術振興事業団 (CREST) の援助を受けて行われたものである。

参考文献

[1] Bregman, A.S. Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press, Cambridge, Mass., 1990.

[2] Bregman, A.S. “Auditory Scene Analysis: hearing in complex environments,” in Thinking in Sounds, (Eds. S. McAdams and E. Bigand), pp. 10–36, Oxford University Press, New York, 1993.

[3] 河原 英紀, “聴覚の計算理論の構築に向けて,” 音響学会誌聴覚研資, H-94-63, Nov. 1994.

[4] Brown, G.J. “Computational Auditory Scene Analysis: A Representational Approach,” Ph. D. Thesis, University of Sheffield, 1992.

[5] Cooke, M. P. “Modeling Auditory Processing and Organization,” Ph. D. Thesis, University of Sheffield, 1991 (Cambridge University Press, Cambridge, 1993).

[6] de Cheveigné, A. “Concurrent vowel identification III: A neural model of harmonic interference cancellation,” J. Acoust. Soc. Am. 101, 2857–2865, 1997.

[7] 柏野 邦夫, “計算機による聴覚の情景解析 –はじめの一歩–,” 音響誌 Vol. 50, no. 12, pp. 1023–1028, Dec. 1994.

[8] Ellis, D. P. W. “Prediction-driven computational auditory scene analysis,” Ph.D thesis, MIT Media Lab., 1996.

[9] Nakatani, T., Okuno, H. G., and Kawabata, T., “Residue-driven Architecture for Computational Auditory Scene Analysis,” In Proc. of IJCAI-95, pp. 165–172, August 1995.

[10] 鵜木 祐史、赤木 正人, “雑音が付加された波形からの信号波形の一抽出法,” 信学論 (A) , vol. J80-A, No. 3, pp. 444-453, March 1997.

[11] Unoki, M. and Akagi, M. “Signal Extraction from Noisy signal based on Auditory Scene Analysis,” In Proc. ICSLP98, Dec. 1998.

[12] McAulary, R. J. and Quatieri, T. F. Low-Rate Speech Coding Based on the Sinusoidal Model, Advanced in Speech Signal Processing. Ed. Furui, S., pp. 165–208, Dekker.

[13] 柏野 邦夫, 田中 英彦, “二つの周波数成分の分離近くに関する工学的モデル,” 信学論 (A) Vol. J77-A No. 5, pp. 731–740, May 1994.

[14] 中野 道雄, 西山 清, パソコンで解くカルマンフィルタ, 丸善, 1993.

[15] 桜井 明, スプライン補間入門, 東京電機大学出版局. 1981.

[16] 武田 一哉, 匂坂 芳典, 片桐 滋, 阿部 匡信, 桑原 尚夫, 研究用日本語音声データベース利用解説書, ATR Technical Report TR-I-0028, 1988.