聴覚の情景解析に基づいた雑音下の定常母音の分離抽出 *

◎ 鵜木　祐史　　赤木　正人 (北陸先端科学技術大学院大学)

1. はじめに

著者らは、Bregman によって提唱された四つの発見的規則 [1] を用いて、人工的な調波複合音と雑音を分離する二波形分離モデルを提案した [2]。しかし、本モデルを音声の分離問題に拡張する際、(1) 分離精度が基本周波数の推定に依存する、(2) 波形分離部における仮定（制約条件）が強い、という課題があった。本稿では、上記２点の改善を行い、雑音中の定常母音を分離抽出する問題の解法を提案する。

2. 二波形分離モデル

著者らが提案した二波形分離モデル [2] は、分離抽出したい信号 $f_1(t)$ を調波複合音、$f_2(t)$ を雑音と仮定し、これらが加算された混合信号 $f(t) = f_1(t) + f_2(t)$ からそれぞれの信号を分離抽出する問題を対象とする。本モデルの信号処理の概要を図 1に、本モデルで利用された制約条件を表 1に示す。

はじめに、混合信号 $f(t)$ のみが観測され（図 1. A)、分析フィルタ群により、振幅包絡 $S_k(t)$ と出力位相 $\phi_k(t)$ に分解される（図 1. B,C)。これは、k 番目の分析フィルタ出力 $X_k(t)$ として表される。

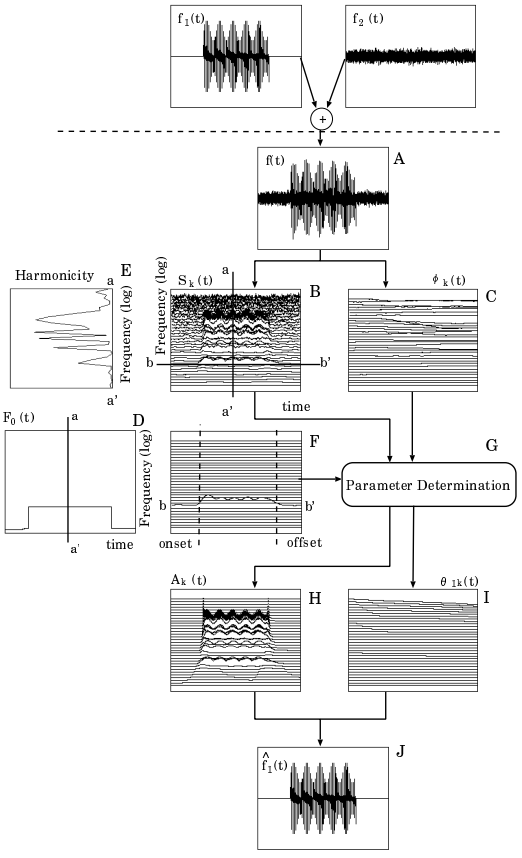

図 1. 二波形分離モデルの信号処理の概要

表 1. 四つの発見的規則と本モデルで利用した制約条件

発見的規則	制約条件
(i) 漸近的変化	時間変化の R 次多項式表現
（ゆっくりと）	（Kalman filtering）
（なめらかに）	（Spline 補間）
(ii) 一つの音響事象に生じる変化	振幅包絡間 $A_k(t)$ の相関
(iii) 調波関係	$F_0(t)$ の整数倍の関係
(iv) 共通の立上り・立下り	チャネル間の立上り・立下りの同期（不一致度）

$$X_k(t) = S_k(t)\sin(\omega_k t + \phi_k(t)) \tag{1}$$

次に、$f_1(t)$ の基本周波数 $F_0(t)$ を求め（図 1. D)、(iii) 調波関係と (iv) 共通の立上り・立下りの制約条件を用いて二波形分離の対象になる時間–周波数領域を決定する（図 1. E,F)。

次に、波形分離部では、観測された $S_k(t)$ と $\phi_k(t)$ から二波形の振幅包絡と入力位相を求める（図 1. G)。ここで、$f_1(t)$ と $f_2(t)$ のフィルタ通過成分をそれぞれ、$A_k(t)\sin(\omega_k t + \theta_{1k}(t))$, $B_k(t)\sin(\omega_k t + \theta_{2k}(t))$ と仮定すれば、二波形の振幅包絡をそれぞれ

$$A_k(t) = S_k(t)\sin(\theta_{2k}(t) - \phi_k(t))/\sin\theta_k(t) \tag{2}$$

$$B_k(t) = S_k(t)\sin(\phi_k(t) - \theta_{1k}(t))/\sin\theta_k(t) \tag{3}$$

で決定できる。但し、$\theta_k(t) = \theta_{2k}(t) - \theta_{1k}(t)$ である。ここで 制約条件 (i) の漸近的変化から、$dA_k(t)/dt = C_{k,R}(t)$ と考えることで、$\theta_k(t)$ を

$$\theta_k(t) = \arctan\left(\frac{S_k(t)\sin(\phi_k(t) - \theta_{1k}(t))}{S_k(t)\cos(\phi_k(t) - \theta_{1k}(t)) + C_k(t)}\right) \tag{4}$$

で決定できる。但し、$C_k(t) = \int C_{k,R}(t) + C_{k,0}$ であり、$C_{k,R}(t)$ は区分的 R 次多項式である。しかし、ここでは、$\theta_{1k}(t)$ と $\theta_{2k}(t)$ を決定する式が不足しているため、これらを一意に決定できない。

そこで、$\theta_{1k}(t) = 0$ と仮定し、$C_k(t) = C_{k,1}(t)$ を、(1) Kalman filter を用いて $C_k(t)$ の推定範囲を決定し、(2) Spline 補間を用いてなめらかな $A_k(t)$ の候補を求め、(3) 振幅包絡 $A_k(t)$ 間の相関値最大を尺度として、最適な $C_{k,1}(t)$ を推定する。この結果から、$\theta_{2k}(t) = \theta_k(t)$ を求め、式 (2) と式 (3) から $A_k(t)$ と $B_k(t)$ を一意に決定する（図 1. H,I)。

最後に、グルーピング部では、調波関係と立上り・立下りの制約条件によって決定された時間周波数領域に対して、波形分離部を実行させる。そして、分

* An Extraction of vowel in noise background based on Auditory Scene Analysis.
By Masashi Unoki and Masato Akagi (JAIST)

離抽出された二波形の成分をそれぞれ一つにまとめ、$\hat{f}_1(t)$ と $\hat{f}_2(t)$ を再構成する（図 1. J)。

3. 二波形分離モデルの改良点

前節で述べた二波形分離モデルを実音声と雑音の分離問題に拡張するため、次の二つの改善を行う。

3.1 分析フィルタ群における基本周波数の推定

本稿では、基本周波数 $F_0(t)$ の推定方法として、比較的雑音にロバストで、分析フィルタ群で推定可能な Comb filtering の方法を用いる。

はじめに、次のような Comb filter を定義する。

$$\mathrm{Comb}(k,\ell) = \begin{cases} 2, & \omega_k = n\cdot\omega_\ell, 1 \le n \le 3 \\ 1, & \omega_k = n\cdot\omega_\ell, 4 \le n \le N \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

ここで、k,ℓ はチャネル番号であり、N は最大高調波次数とする。次に、時刻 t における Comb filter の通過量を求め、これを最大にする $\hat{\ell}$ を求める。

$$\hat{\ell} = \arg\max_{\ell \le L} \sum_{k=1}^{K} \mathrm{Comb}(k,\ell) S_k(t) \quad (6)$$

但し、L は ℓ の探索範囲の上限である。この $\hat{\ell}$ に対応する $X_k(t)$ の中心周波数を基本周波数 $F_0(t) = \omega_{\hat{\ell}}/2\pi$ とする。本稿では、$N = 10$、$L = K/4$ とした。

3.2 $\theta_{1k}(t)$ の制約条件の緩和

先の論文 [2] では、制約条件 (i) の漸近的変化を $\theta_{1k}(t) = 0$ とした。これは、人工的な調波複合音などの分離抽出ではほとんど問題とならないが、実音声を考慮する場合、とても強い条件となる。そこで、この制約条件を緩和するために、$d\theta_{1k}(t)/dt = D_{k,R}(t)$ と考える。ここで、$D_{k,R}(t)$ は区分的 R 次多項式である。本稿では、$D_{k,R}(t) = 0$、つまり $\theta_{1k}(t) = D_{k,0}$ と考え、$-\pi/2 \le D_{k,0} \le \pi/2$ の範囲内で、振幅包絡 $A_k(t)$ 間の相関値最大を尺度として決定する。この結果、$C_{k,1}(t)$ により $\theta_k(t)$ が決定され、$D_{k,0}$ の結果から、$\theta_{1k}(t)$, $\theta_{2k}(t)$ が決定される。また、式 (2) と式 (3) から $A_k(t)$ と $B_k(t)$ が決定される。

4. 二波形分離のシミュレーション

本稿では、上記の改善により、二波形分離モデルが雑音下から実母音を分離抽出できることを確認するために、次のような計算機シミュレーションを行った。データは、$f_1(t)$ を ATR datebase mau 氏の /a/（図 2(a)）、$f_2(t)$ をピンク雑音とし、二つの波形の SNR が $10, 20, 30$dB となる 3 種類の混合信号 $f(t)$ を用いた。また、分離精度については、$f_1(t)$（真値）と分離抽出された $\hat{f}_1(t)$ に対し、フレーム（フレーム長 51.2 ms, シフト 25.6 ms, Hamming 窓）単位に求めた振幅スペクトル歪み [2] の平均値で評価した。

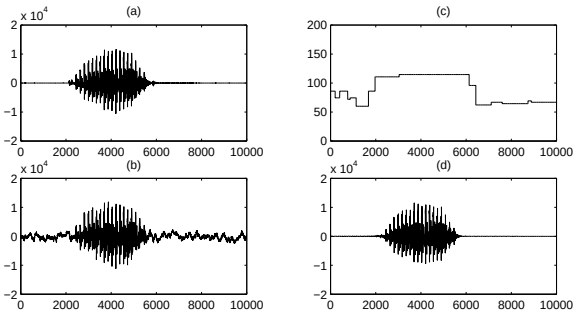

図 2. シミュレーションの例：**(a)** $f_1(t)$ 実母音/a/, **(b)** $f(t)$ **(SNR= 10 dB)**, **(c)** 推定された $F_0(t)$, **(d)** $\hat{f}_1(t)$

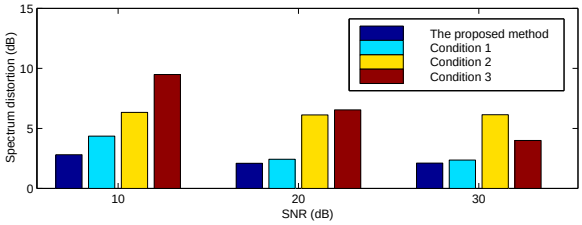

図 3. 分離特性

例えば、SNR が 10 dB の $f(t)$ に対し（図 2(b))、本モデルは図 2(c) に示す $F_0(t)$ を抽出し、図 2(d) に示す $\hat{f}_1(t)$ を分離抽出する。また、(Cond.1) Comb filter による調波成分抽出＋ Kalman filter による $A_k(t)$ の推測、(Cond.2) Comb filter による調波成分抽出、(Cond.3) 処理なし、の 3 つの比較を行ったところ、図 3に示す結果を得た。ここで、(Cond.1) は、表 1 の漸近的変化（なめらかさ）を、(Cond.2) は漸近的変化すべてを、(Cond.3) はすべての制約条件を省略した場合に相当する。この結果から、本方法の有効性が確認できる。また、SNR が 10, 20, 30 dB のとき、それぞれスペクトル歪み量で 6.7, 4.5 ,1.9 dB、雑音を除去できることがわかる。

5. まとめ

本稿では、二波形分離問題の枠組を示し、雑音中の定常母音の分離抽出法を提案した。この方法は、不良設定問題となる二波形分離問題を、Bregman によって提唱された四つの発見的規則を用いて一意に解くものである。本方法の分離例として、雑音中の実母音 /a/ を分離抽出する問題を解いた。また、他の方法と比較することで、本方法の有効性を示した。

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、科学技術振興事業団（CREST)、および科学研究費補助金（10680374）の援助を受けて行われた。

参考文献

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

[2] 鵜木、赤木：“基本周波数の時間変動を考慮した調波複合音の抽出法,” 信学技報, SP97-129, March 1998.