基本周波数の時間変動を考慮した調波複合音の抽出 *

◎ 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) 立上り・立下り、(2) 漸近的変化、(3) 調波関係、(4) 1 つの音響事象に生じる変化) [1] を利用することで、雑音中から AM 調波複合音を分離抽出する方法を提案した [2]。本稿では、この方法において、基本周波数の推定とその時間変動の対応を考慮した分離・抽出方法を提案する。

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は、図 1に示すような (a) 分析フィルタ群、(b) 波形分離部、(c) グルーピング部の 3 ブロックで構成される [2]。

はじめに、混合信号 $f(t)$ ($= f_1(t) + f_2(t)$) のみが観測され、Gammatone filter を基底関数とした wavelet 分析合成系 (チャネル数 $K = 128$、解析周波数範囲 60 ~ 6000Hz) で構成された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 $f_2(t)$ の成分は

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

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

と仮定され、これらの和となる $X_k(t)$ は

$$X_k(t) = S_k(t)\sin(\omega_k t + \phi_k(t)) \tag{3}$$

と表される。

次に、波形分離部では、四つの物理パラメータ ($S_k(t)$, $\phi_k(t)$, $\theta_{1k}(t)$, $\theta_{2k}(t)$) を用いて二波形の振幅包絡 $A_k(t)$ と $B_k(t)$ を次式で一意に決定する [2]。

$$A_k(t) = S_k(t)\sin(\theta_{2k}(t) - \phi_k(t))/\sin\theta_k(t) \tag{4}$$

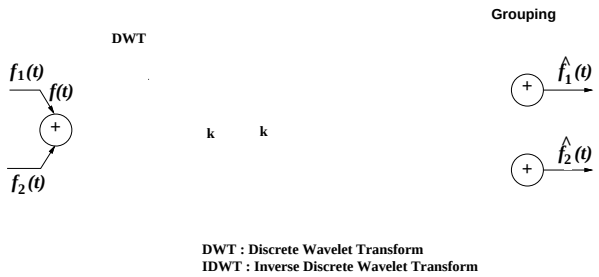
$$B_k(t) = S_k(t)\sin(\phi_k(t) - \theta_{1k}(t))/\sin\theta_k(t) \tag{5}$$

但し、$\theta_k(t) = \theta_{2k}(t) - \theta_{1k}(t)$ である。ここで、$S_k(t)$ と $\phi_k(t)$ はそれぞれ wavelet 変換で定義された振幅、位相スペクトルから決定できる。残る二つの入力位相のうち、一つは $\theta_{1k}(t) = 0$ と仮定され、$\theta_{2k}(t)$ は発見的規則 (2) と (4) を利用して一意に決定される。

最後に、グルーピング部では発見的規則 (1) と (3) を用いて分離された $A_k(t)$ と $B_k(t)$ をグループ化し、$f_1(t)$ と $f_2(t)$ を再構成する。

2.1 $\theta_{2k}(t)$ の決定方法

入力位相 $\theta_{2k}(t)$ は図 2に示す三つの制約条件を用いることで得られる [2]。はじめに、制約条件 (i) に



---

* An Extraction of the Complex tone considering temporal variation of the fundamental frequency.
By Masashi Unoki and Masato Akagi (JAIST)

図 1．二波形分離モデル

(i) ゆっくりと (漸近的変化) **(Bregman)**

$$dA_k(t)/dt = C_{k,R}(t) \quad (6)$$

$$\Rightarrow \quad \theta_k(t) = \tan^{-1}\left(\frac{S_k(t)\sin\phi_k(t)}{S_k(t)\cos\phi_k(t)+C_k(t)}\right) \quad (7)$$

(ii) なめからに (漸近的変化) **(Bregman)**

$$\sigma = \int_{t_a}^{t_b}\left[A_k^{(R+1)}(t)\right]^2 \Rightarrow \min \quad (8)$$

(iii) 1 つの音響事象に生じる変化 **(Bregman)**

$$\frac{A_k(t)}{\|A_k(t)\|} \approx \frac{A_\ell(t)}{\|A_\ell(t)\|}, \quad \ell \neq k \quad (9)$$

$$\Rightarrow \max_{\hat{C}_k - P_k \leq C_k \leq \hat{C}_k + P_k} \frac{< \hat{A}_k, \hat{\hat{A}}_k >}{||\hat{A}_k||||\hat{\hat{A}}_k||} \quad (10)$$

図 2．二波形分離の制約条件 (入力位相 $\theta_{2k}(t)$ の決定)

より、$\theta_{2k}(t)$ の一般解 (7) を得る。次に、これを決定するためには未定関数 $C_k(t)$ を知る必要があるため、Kalman フィルタを用いて推定する。ここで、$\hat{C}_k(t)$ は最小分散推定値であり、$P_k(t)$ は推定誤差である。次に、制約条件 (ii) を満たす $A_k(t)$ を決めるために、推定誤差内で Spline 補間された $C_k(t)$ の候補を求める。最後に、制約条件 (iii) より、分離・抽出したい信号の振幅包絡 $A_k(t)$ 間の相関が最大になる $C_k(t)$ を式 (10) から求めることで、最適な $C_k(t)$ を求め、$\theta_{2k}(t)$ を一意に決定する。

2.2 基本周波数の推定とその時間変動の対応

本モデルでは、河原によって提案された TEMPO[3] を利用して調波複合音の基本周波数を抽出する。ここで、TEMPO で推定された基本周波数を $F_0(t)$ とする。一般に、$F_0(t)$ は時間的に変化するが、その変動は急激には起こらないものと解釈できる。そこで、$F_0(t)$ の時間変化に対し、図 3に示す制約を適用する。これは、ある微小区間における $F_0(t)$ が一定であるとし、$F_0(t)$ の分散量が一定である区間 ($t_h - t_{h-1}$) を微小区間と解釈して決定する。本稿では、$\Delta F_0 = 1$ Hz とした。

(iv) 変化は急激に起こらない (漸近的変化) **(Bregman)**

$$dF_0(t)/dt \quad = 0 \tag{11}$$

$$\Rightarrow \quad \frac{1}{t_h - t_{h-1}} \int_{t_{h-1}}^{t_h} \left| F_0(t) - \overline{F_0(t)} \right|^2 dt \leq \Delta F_0^2 \tag{12}$$

図 **3** . 基本周波数の時間変動に対する制約条件

(v) 調波関係 **(Bregman)**

$$k = \frac{K}{2} - \left\lceil \frac{\log(n \cdot F_0(t)/f_0)}{\log \alpha} \right\rceil, \qquad n = 1, 2, \cdots, N \tag{13}$$

(vi) 共通の立上り・立下り **(Bregman)**

$$|T_{\rm S} - T_{k,\rm on}| \leq 50\text{msec}, \quad |T_{\rm E} - T_{k,\rm off}| \leq 100\text{msec} \tag{14}$$

図 **4** . グルーピングの制約条件

2.3 グルーピングの制約

グルーピング部は、制約条件 (iv) で得られた各微小区間に対して図 4に示す制約条件を満たすとき、$X_k(t)$ から二波形の振幅包絡 ($A_k(t)$, $B_k(t)$) を求め、それらをグルーピングして信号波形に再構成する。ここで、制約条件 (v) は $f_1(t)$ の基本周波数 $F_0(t)$ に対する高調波成分の分離抽出を制約するもの、制約条件 (vi) は、$f_1(t)$ の非高調波成分を分離抽出するために基本波の立上りと立下りが同期する非高調波成分を制約するものに位置付けられる。また、制約条件 (vi) における $T_{k,\rm on}$ と $T_{k,\rm off}$ は $X_k(t)$ における高調波成分の立上りと立下りであり、$dS_k(t)/dt$ と $d\phi_k(t)/dt$ から検出する。但し、$f_1(t)$ の基本波の立上りを $T_{\rm S} = t_{h-1}$、立下りを $T_{\rm E} = t_h$ とする。

3. 二波形分離のシミュレーション

本シミュレーションでは、$f_1(t)$ を図 5に示す LMA 合成母音/a/、$f_2(t)$ をピンク雑音とする。但し、$F_0(t)$ の平均は 125 Hz であり、その下限・上限が 123 ~ 128 Hz である。また、混合信号 $f(t)$ は SNR が $0, 5, 10, 15, 20$dB となる 5 種類の信号とし、フレーム (フレーム長 51.2msec、フレームシフト 25.6msec、Hamming 窓) 単位に求めた SD (Spectrum Distortion) の平均値で分離精度を評価した。

例えば、図 6に示すように $f(t)$ の SNR が 10 dB のとき、本モデルは図 7に示すような $\hat{f}_1(t)$ を分離抽出できる。このとき、$\hat{f}_1(t)$ の振幅スペクトル (SD と同条件) は、図 8に示すように明らかに雑音成分が除去されている。また、図 9に示す分離精度の結果において、$f(t)$ と $\hat{f}_1(t)$ の SD 値の差から、本方法は約 15 dB 程度雑音を除去でき、雑音中から望みの調波複合音を分離抽出できることがわかる。

4. まとめ

本稿では、著者らが提案した AM 調波複合音の分離・抽出法において、基本周波数の抽出および基本周波数の時間変動の対応を考慮することで、雑音中から調波複合音を分離抽出する方法を提案した。調波複合音として LAM 合成母音を利用したところ、本方法により、雑音が付加された音声から望みの音声を分離抽出することが可能になった。このとき、SD 値で約 15 dB の雑音除去が実現できた。

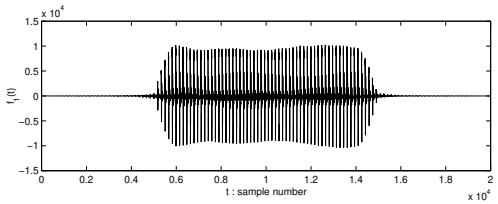


図 **5** . **LMA** 合成母音/**a**/ $f_1(t)$

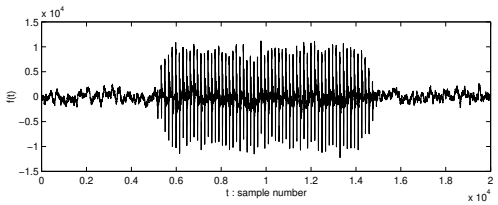


図 **6** . 混合信号 $f(t)$ (**SNR**= 10 **dB**)

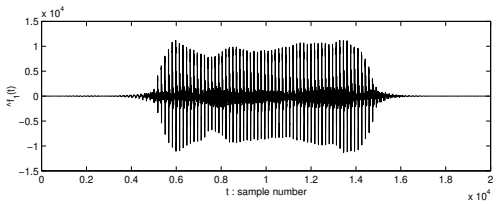


図 **7** . 分離抽出された信号 $\hat{f}_1(t)$ (**SNR**= 10 **dB**)

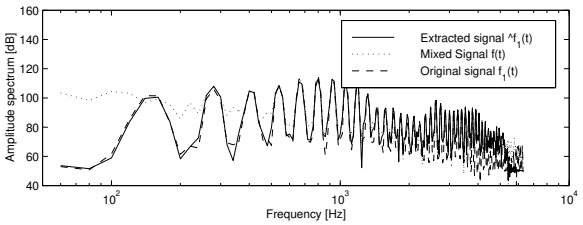


図 **8** . 振幅スペクトルの比較 (**SNR**= 10 **dB**)

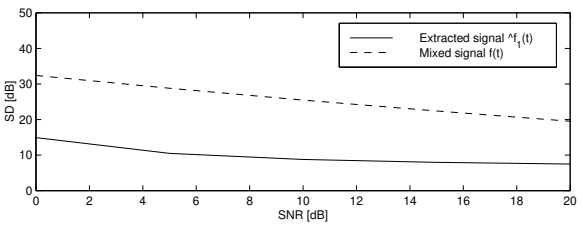


図 **9** . 分離抽出された信号 $\hat{f}_1(t)$ の **SD** 特性

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] Masashi Unoki, Masato Akagi, "A Method of Signal Extraction from Noisy Signal," In Proc. EuroSpeech'97, vol. 5, pp. 2583-2586, RHODOS-GREECE, Sept. 1997.

[3] Hideki Kawahara, "STRAIGHT – TEMPO: A Universal Tool to Manipulate Linguistic and Para-Linguistic Speech Infomation," In Proc. SMC-97, Oct. 12-15, Orland, Florida, USA.