Vol.2018-MUS-119 No.29
Vol.2018-SLP-122 No.29
2018/6/16

# 準周期性を考慮した複合ウェーブレットボコーダによる音声分析合成

小口 純矢[1,a)]　嵯峨山 茂樹[1,b)]

**概要**：本論文は, 波形生成モデルである複合ウェーブレットボコーダのさらなる品質の向上を目指す. これまでに我々は, 音声分析合成や HMM 音声合成系において, 複合ウェーブレットボコーダの安定性を示してきた. ここで、さらに WORLD や STRAIGHT で用いられている非周期性指標のような, 音声の準周期性を取り入れることができれば, 有声摩擦音やかすれ声のように周期成分と非周期成分の両方を持つ音声を表現でき, 高品質な音声を合成できると期待される. 本論文では, 複合ウェーブレットの基本波形を完全な周期ではなく Jitter を付与した準周期的なパルス列によって駆動させることで実現した. また, 主観評価実験により, 改良後の音声が改良前の音声より品質が有意に高いことを示した.

**キーワード**：音声分析合成，準周期性, Wavelet，GMM

# Analysis/Synthesis using Composite Wavelet Vocoder Considering Quasi-periodicity

Junya Koguchi[1,a)]　Shigeki Sagayama[1,b)]

## 1. はじめに

高品質かつ加工性の高い統計的音声合成システムを実現する上では, 音響モデルだけでなく, 統計学習によって推定される音響特徴量から音声波形を合成するボコーダの品質は合成音声の品質を大きく左右する. そのため音声合成系おいて, 音声の持つ本質的な特徴を捉えたパラメータを抽出し, また抽出したパラメータから高品質な音声を合成し, さらには人工的なパラメータの変更に対して頑健であることがボコーダには要求される.

近年では, End-to-End モデルと呼ばれる音声波形をそのまま学習し, 音声波形を直接出力する方式 [1] を用いたボコーダレスなシステムもいくつか提案されている. しかし, End-to-End モデルは学習データに含まれないパラメータを入力した際に品質が劣化することや, ボコーダ方式はユーザが合成後の音声を好みに合わせて加工できるという点から, ボコーダを用いた音声合成は依然として有用である.

音声合成系に用いられるボコーダには, LPC ボコーダ [2] をはじめ, Griffin・Lim アルゴリズム [3] を用いてスペクトルから直接音声波形を得る方式 [4] など様々なモデルが提案されている. 中でも STRAIGHT[5] や WORLD[6](D4C edition[7]) は高品質な音声分析合成系として広く用いられている.

一方で、我々は複合ウェーブレットボコーダ [8], [9] を安定した音声を合成できるボコーダとして音声合成系に利用してきた. なお, これまでの研究においてスペクトルの特徴量表現のモデルとして, 複合ウェーブレットモデル (Composite Wavelet Model; CWM) と呼んでいたが, ここではボコーダとしての側面を検討するため, ここでは複合ウェーブレットボコーダ (Composite Wavelet Vocoder; CWV) と呼ぶことにする. CWV による音声波形合成は FIR フィルタによる畳み込みと解釈でき, インパルス応答

[1] 明治大学
Meiji University, Nakano, Tokyo 164–8525, Japan
[a)] ev50552@meiji.ac.jp
[b)] sagayama@meiji.ac.jp

Vol.2018-MUS-119 No.29
Vol.2018-SLP-122 No.29
2018/6/16

が短く, 急激なパラメータ変動に対しても品質劣化が起こりにくいことが報告されている [8].

さらに, 混合励振源モデル [10] のように, 周期成分と非周期成分の両方を含む音声を表現するため, 有声区間と無声区間とを連続的に扱うことで品質の向上が期待できる. そこで本研究では, 音声の周期性のゆらぎに注目し, 複合ウェーブレットの基本波形を完全な周期ではなく Jitter を付与した準周期的なパルス列によって駆動させる手法を取り入れた結果を報告する.

## 2. CWV 音声分析合成系

### 2.1 音響特徴量抽出

CWV を用いた音声分析及び分析によって得られたパラメータからの波形生成について述べる [8], [9]. CWV では, 音声スペクトル包絡を GMM によって近似し, 得られた平均, 分散, 重みをスペクトル包絡の特徴量表現 (以後, CWV パラメータと呼ぶ) とみなす. 各ガウス関数の平均はスペクトルのピークに対応すると解釈でき, ピークの周波数変動を記述する上で都合が良い. この時, 補助関数法によってスペクトル包絡と GMM によるモデルスペクトルとの間の $I$-divergence を逐次的に最小化することで, CWV パラメータによるスペクトル包絡の近似を行う (Fig. 1). ここで $I$-divergence とは,

$$I(Y|F) = \sum_{\omega,t}[Y_{\omega,t}\log\frac{Y_{\omega,t}}{F_{\omega,t}} - Y_{\omega,t} + F_{\omega,t}] \tag{1}$$

と定義される. ただし $Y_{\omega,t}, F_{\omega,t}$ は, それぞれ時刻 $t$ における観測およびモデルスペクトル包絡を表す. $F_{\omega,t}$ は以下のような GMM で表現される.

$$F_{\omega,t} = \sum_{k=1}^{K}\frac{w_k}{\sqrt{2\pi\sigma_k^2}}\exp\left[-\frac{(\omega-\mu_k)^2}{2\sigma_k^2}\right] \tag{2}$$

$K$ は GMM の混合数を表す. 最終的に得られた CMV パラメータ $\mu_k, \sigma_k, w_k (k = 1, \ldots, K)$ を連結し, これをスペクトル特徴量として用いる. なお, 本研究では観測スペクトルを WORLD における CheapTrick[11] を用いて推定されたスペクトル包絡とし, GMM 推定の際の $\mu_k$ の初期値として, それぞれ $2K$ 次の LSP 解析によって得られたスペクトル対の平均を与え, 分散 $\sigma_k$ は 10 を与えた.

### 2.2 時間遷移確率モデルの導入

前セクションで述べた CWV パラメータの抽出法は, フレーム毎に独立に抽出を行うため, このままでは得られた特徴量のインデックスの交代が起きる. 各混合成分の平均はフォルマント周波数に, 分散がフォルマントの広がりに対応すると解釈できるため, 時刻フレーム毎の処理では, 存在するはずのスペクトルのピークを近似し損ねたり, フォルマントの滑らかな軌跡を捉えることができないといった問題が起こる. そこで, CWV パラメータにおける $\mu_k$ の時

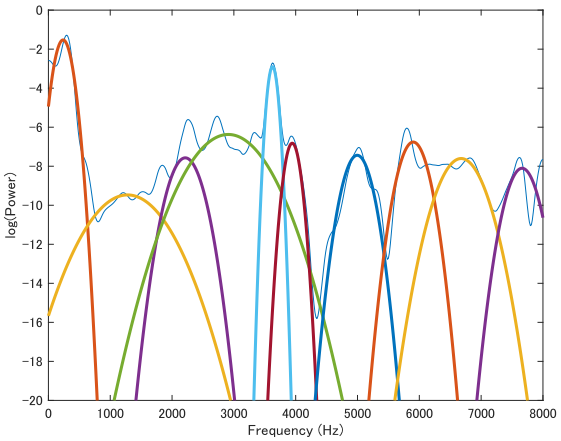


**Figure 1** 音素/i/におけるスペクトル包絡の GMM 近似;($K$=10).

間遷移確率を導入する. 時刻 $t$ における $\mu_k$ を $\mu_k^{(t)}$ としたとき, $\mu_k^{(t)}$ の時間変動は, 平均を 1 つ前のフレームの $\mu_k^{(t-1)}$ の正規分布に従うという仮定をおく (3).

$$P(\mu_k^{(t)}|\mu_k^{(t-1)}) = \mathcal{N}(\mu_k^{(t)}; \mu_k^{(t-1)}, \nu_k^2) \tag{3}$$

ここで, 各インデックス $k$ おける分散 $\nu_k^2$ は, 許容される $\mu_k$ の時間変動の程度を表している. この時間遷移確率の導入は, 抽出される平均値パラメータを時間方向に平滑化し, $mu_k$ 系列の時間的な連続性を保つ効果がある. 遷移確率モデルの導入により, CWV パラメータの抽出は観測スペクトログラムとモデルスペクトログラムとの違いを表す以下の目的関数の最小化によって実現される.

$$J(Y|F) = I(Y|F) + \sum_{k}^{K}\frac{1}{2\nu_k^2}\sum_{t=1}^{T-1}(\mu_k^{(t+1)} - \mu_k^{(t)})^2 \tag{4}$$

ただし $T$ は音声から得られたスペクトルの全フレーム数である. [12] と同様に, Jensen の不等式を用いて補助関数を導入する. このとき, $\mu_k$ を除くパラメータは, 2.1 における式 (2) によって更新する. 一方, $\mu_k$ については, 時間遷移確率モデルを考慮した以下の式により更新される. 時系列ベクトルを $\boldsymbol{\mu_k} = (\mu_k^{(1)}, \mu_k^{(2)}, \cdots, \mu_k^{(T)})^{\mathrm{T}}$ とすると, $\{\mu_k^{(t)}\}$ の更新式は,

$$\boldsymbol{\mu_k^*} = \frac{1}{2}(D_k + E_k)^{-1}F_k \tag{5}$$

と書ける. ただし,

$$D_k = \frac{1}{2\nu_k^2}\left\{D^{(i,j)}\right\}_{i,j} \quad (i,j \in [1,T]) \tag{6}$$

$$D^{(i,j)} = \begin{cases} 1 & (i=j=\{1,T\}) \\ 2 & (i=j\in[2,T-1]) \\ -1 & (|i-j|=1) \\ 0 & (\text{other}) \end{cases} \tag{7}$$

$$E_k = \frac{1}{2\sigma_k^2}\left\{E_k^{(i,j)}\right\}_{i,j} \quad (i,j \in [1,T]) \tag{8}$$

$$E_k^{(i,j)} = \begin{cases} \sum_\omega Y(\omega,i)\lambda_k(\omega,i) & (i=j) \\ 0 & (i\neq j) \end{cases} \quad (9)$$

$$F_k = \frac{1}{\sigma_k^2}(F_k^{(i)})_i \quad (i\in[1,T]) \quad (10)$$

$$F_k^{(i)} = \sum_\omega \omega Y(\omega,i)\lambda_k(\omega,i) \quad (11)$$

である．ただし, $\lambda_k$ は補助関数法における補助変数である．

時間遷移確率モデルを導入しない場合の特徴量抽出の結果を Fig. 2 に，時間遷移確率モデルを導入した場合の結果を Fig. 3 に示した．用いた音声，GMM の混合数，初期値の設定は同一である．$\nu_k$ の値については，$\mu_k$ の時間変化の標準偏差がメル周波数軸上でほぼ一定となるよう定めた．Chain を導入しなかった場合にしばしば確認されていたインデックスの交替が, $\mu_k$ の時間変動が平滑化されたことにより，連続的に変動していることが確認できる．

### 2.3 CWV パラメータからの音声波形合成

CWM 特徴量と基本周波数情報を用いて，音声波形を合成する手法について述べる．1 章で挙げた巡回型フィルタの問題点を解決する方法として，GMM 包絡近似の逆フーリエ変換から得られる FIR 型フィルタを用いる．ここで, 式 (12) で示されるようにガウス関数の逆フーリエ変換が Gabor 関数，つまりガウス関数と三角関数の積であることに注目する．この性質を利用し, GMM に対して逆フーリエ変換を施すことで, Gabor wavelet の基本波形を得ることができる.[8].

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(\omega-\mu)^2}{2\sigma^2}\right] \Leftrightarrow \frac{1}{\sqrt{2\pi^2}}\exp\left[-\frac{\sigma^2t^2}{2}+j\mu t\right] \quad (12)$$

有声音部分では, 時間領域において Gabor wavelet の基本波形を基本周波数に対応する周期の間隔で並べることにより音声波形を得る．これは FIR フィルタを基本周波数に対応したインパルス列で駆動することと等価である．無声音部分の合成には, 波形の非周期性が高くなることを考慮し, 基本波形の並べる間隔をランダムにすることによって実現する．しかし, この手法では有声音と無声音とで合成手法が区別されており, かすれ声や有声摩擦音のような周期性と非周期性を持つ音声を十分に表現できない．

## 3. 準周期性を考慮した CWV 音声波形合成

2 章で述べたように, 有声音の場合と無声音の場合で Wavelet の設計手法が異なる問題を解決するため, 周期性と非周期性を連続的に扱えるモデルを導入する必要がある．ここで, 我々は声帯振動周期のゆらぎに注目する．有声音の合成における CWV 基本波形の周期的な並びを徐々にランダムにしていくことによって, 周期成分の高い音声から非周期性の高い音声へと連続的に変動させることができる．本研究では, 1 次元の非周期性指標と Jitter には線形の対応があるという仮定のもと, 基本波形を最小 0ms, 最大

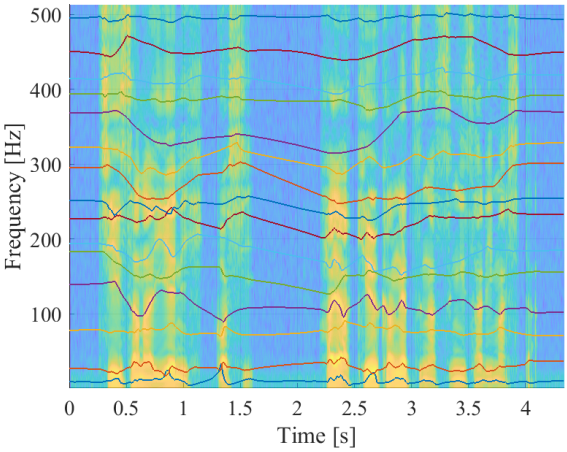


**Figure 2** Chain を導入しない場合の $\mu_k$ 抽出結果 ($K$=10).

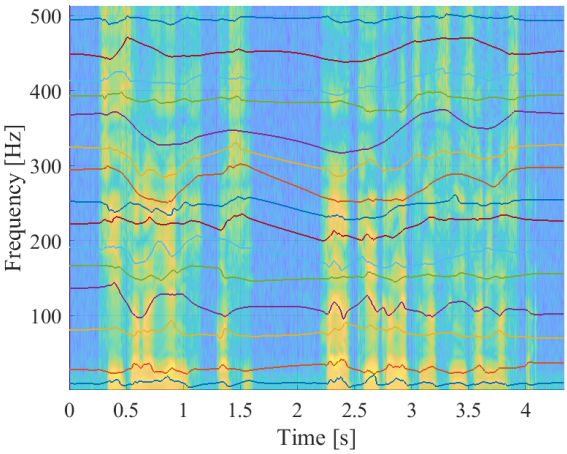


**Figure 3** Chain を導入した場合の $\mu_k$ 抽出結果 ($K$=10).

±3ms の Jitter を一様乱数によって付与した間隔で並べて合成を行った．分析合成時には, まず WORLD の D4C?によって 1 次元の非周期性指標の抽出を行った後, それに対応する乱数のパラメータを用いて音声波形の合成を行う．

## 4. 実験的評価

### 4.1 実験条件

改良した CWV の品質を比較するため, 自然音声をリファレンス音声とする Degradation MOS (DMOS) テストを実施した．Jitter を付与する改良を施していない CWV による合成音声 (CWV) と改良を施した CWV による合成音声 (CWV+J), WORLD による合成音声の品質 (WORLD) を同時に評価した．リファレンスと分析合成に用いた音声は ATR 日本語話者データベースの中から, 男性話者によって読み上げられた音素バランス 503 文のうち無作為に選んだ 5 音声を使用した．音声のサンプリングレートは 16kHz, CWV における GMM の混合数は改良前と後でともに 25 とした．聴取実験には 10 代から 20 代の男女 10 人が被験者として参加し, 各々が普段使用しているヘッドホンまたはイヤホンを用いて静かな室内で行われた．

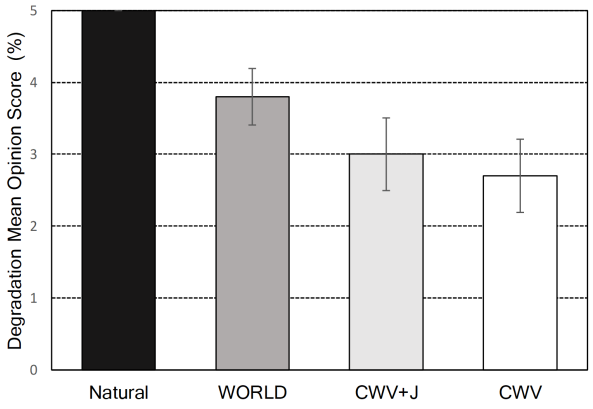


**Figure 4** 主観評価実験結果.

### 4.2 実験結果と考察

結果を Fig. 4 に示す. 改良を施していない音声よりも改良後の音声のスコアが高いことから, Jitter の付与は有効に働いたことが確認できる. しかし, WORLD の合成音声が最も高いスコアを示しており, CWV は WORLD が実現する品質には達成できていないことが分かる. これは零位相化の影響で Wavelet の基本波形の中心にパワーが集中するため, ブザーのようなノイズが生じてしまうことに起因すると考えられる.

## 5. おわりに

本研究では，巡回型フィルタの時間特性および利得特性の問題を解決する安定性の高い音声分析合成系である CWV に対し, 準周期なパルス列によって駆動させるモデルを組み込み, 周期成分と非周期成分の両方を持つ音声を連続的に表現できる手法を提案した. 主観評価実験の結果, この手法により CWV の合成音声品質の改善を確認した. しかし, 非周期性指標と Jitter 成分との理論的な関係が不明瞭であり, ノイズの軽減も課題である. 合成手法の改良や DNN 音声/歌声合成系への組み込み, CWV パラメータにおける平均の時系列の変動に対して変調スペクトルのようなモデルを導入するなど, 拡張に取り組む予定である.

**謝辞** 本研究は科研費基盤A (課題番号 17H00749) の支援を受けて行われた.

### 参考文献

[1] Y. Wang, et al. Tacotron: Towards end-to-end speech synthesis. *https://arxiv.org/abs/1703.10135*, 2017.
[2] F.Itakura and S. Saito. Analysis synthesis telephony based on the maximum likelihood method. *Journal of the Royal Statistical Society B*, Vol. 39, pp. 185–197, 1968.
[3] D. Griffin and J. Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 32, No. 2, pp. 236–243, 1984.
[4] Y. Hamada, et al. Non-filter waveform generation from cepstrum using spectral phase reconstruction. *Proc. SSW*, pp. 28–32, 2016.
[5] H. Kawahara, et al. Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based f0 extraction. *Speech Communication*, Vol. 27, No. 3, pp. 187–207, 1999.
[6] M. Morise, et al. World: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, Vol. E99.D, No. 7, pp. 1877–1884, 2016.
[7] M. Morise. D4c, a band-aperiodicity estimator for high-quality speech synthesis. *Speech Communication*, Vol. 84, pp. 57–65, 2016.
[8] 槐ほか. 複合ウェーブレットモデルに基づく音声の分析合成. 電子情報通信学会技術研究報告. SP, 音声, Vol. 105, No. 370, pp. 1–6, 2005.
[9] 北条ほか. 複合ウェーブレット分析合成系に基づく hmm 音声合成. 日本音響学会研究発表会講演論文集 (CD-ROM), Vol. 2012, pp. ROMBUNNO.2–2–7, 2012.
[10] A. V. McCree and T. P. Barnwell. A mixed excitation lpc vocoder model for low bit rate speech coding. *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 4, pp. 242–250, 1995.
[11] Cheaptrick, a spectral envelope estimator for high-quality speech synthesis. *Speech Communication*, Vol. 67, pp. 1–7, 2015.
[12] 亀岡. スペクトル包絡と調波構造の合成関数モデルによる音声分析. 日本音響学会 2005 年秋季研究発表会講演論文集, No. 2-6-7, 2005.