複合ウェーブレットモデルによる音声合成の検討*

◎槻武也† 松本恭輔‡ 酒向慎司† 嵯峨山茂樹† （† 東大・情報理工, ‡ 東大・工）

1 はじめに

筆者らは、テキスト音声合成などへの応用を目的とした新しいパラメトリックな音声分析合成モデルとして、複合ウェーブレットモデル (CompositeWavelet Model、略称 CWM) 法を提案した [1]。テキストの読み上げにとどまらず多様な発話スタイルの音声を合成するための方針として、柔軟なパラメータの操作による加工があげられる。パラメトリックな音声分析合成系による HMM 音声合成では多様な発話スタイルの実現が期待されているものの、合成音声の品質が十分ではない。本稿ではその原因としてフィルタ手法の時間特性に関する問題について議論し、それが CWM 法で改善されることを分析合成音における時間特性の比較によって示す。

2 従来の音声合成法の問題点

2.1 波形接続方式とフィルタ方式

波形接続型の音声合成は、十分なバラエティの音声素片のデータがあれば高品質な音声合成が期待できるが、ノンパラメトリックな手法であるため、データベースに含まれない条件の音声を生成するような加工性は高くない。また、多様な音声データを収集する必要があり、感情音声や対話音声などを生成するには効率が悪いと考えられる。

一方、パラメトリックな音声合成手法の代表例はフィルタ型の音声合成であり、スペクトル包絡と微細構造を (近似的に) 分離して扱う。そのため、ピッチ周波数を任意に操作でき、フィルタ特性を比較的少数のパラメータで表現して音声スペクトルを生成するため、加工性が高いと期待されている。フィルタ特性を与えるパラメータとして LPC、PARCOR、LSP やケプストラムなどを用いる場合は、音声分析合成方式が確立されている。

2.2 フィルタ方式音声合成の高 Q 値問題

フィルタ型の音声合成はピッチ周波数の操作が可能であるが、駆動する基本周波数によってはエコーのかかっているような印象の、歯切れの悪い音声が合成されることがある。その一因として、われわれはフィルタの利得特性と時間特性に注目した。

全極型フィルタによる音声分析合成方式 (LPC 系) における有声音の分析合成について考察する。音声スペクトル包絡の山と谷の間には大きなレベル差があることが多く、これを少数のパラメータを用いたモデルで表現するために、比較的次数が低い全極型フィルタを用いる。全極型フィルタは多重共振系であるが、このような理由でおのおのの極の共振特性の Q 値は、実際の声道の特性よりも大きな値を取る傾向がある。このような周波数特性のフィルタの時間特性は、共振周波数の信号成分に対して Q 値に比例した利得が生じるとともに、Q 値に比例した時定数で出力振幅が立上り、減衰する。

このため、分析時と異なる基本周波数で全極型フィルタを駆動すると、倍音成分が高 Q 値の共振周波数に一致した場合などには利得がゆっくり増大するとともに時間特性が悪くなり、「歯切れの悪い」音になると考えられる。それを改善するために Q 値を下げると、包絡の山と谷のレベル差が形成できず buzzy な音が生成される。このように、有声音のピッチ周期によって利得変動が大きく、合成音声出力振幅を制御できない問題は有極型フィルタ (巡回型ディジタルフィルタ) の本質に根ざす問題で解消は難しい。極零モデルはやや有望であるが、分析合成法として確立していない。また、フィルタでは出力信号の利得が Q 値に比例するため、その利得は駆動音原信号のピッチ周波数によって大きく変動する。このため、フィルタ型音声合成では合成音声のパワーの制御が難しい。

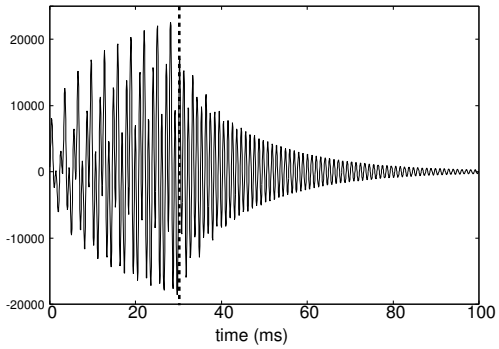

図 1: LPC フィルタ出力の時間特性の例

図 1 は、音素/o/を LPC 分析して得たフィルタに 30ms の長さのインパルス列を入力したときの出力波形である。入力に対して出力振幅は増大を続ける (定常状態に達するまでに時間が掛かる) とともに、入力終了後も数十 mS にわたり出力が持続している。

2.3 複合正弦波モデル (Composite Sinusoidal Model) 音声合成

CSM 音声合成 [2] では、線スペクトルモデルに基づく音声分析法である CSM 音声分析によって、フォルマント周波数に対応する複数個の正弦波周波数 (CSM 周波数) を得る。そして、それらの周波数の正弦波の和において、位相を基本周期ごとに 0 にリセットすることにより有声音を合成する。これは、巡回型フィルタを用いないパラメトリックな音声合成方式であり、振幅の制御は極めて容易であるため「歯切れのよい」音声合成が期待できるが、音声スペクトルの近似方法としては検討の余地があった。

3 複合ウェーブレットモデル (CWM)

3.1 波形小片の接続による音声合成

前説で述べた方式における有声音の合成を、ピッチ周期のインパルス列を入力したある線形系と考える。これらを線形系のインパルス応答により整理すると、PSOLA 方式あるいは波形接続型では音声波形のピッチ周期波形そのものをインパルス応答とすることに対し、全極型フィルタでは推定されたスペクトル包絡の逆 Fourier 変換が対応する。

これを基本波形の繰り返しとして解釈し比較すると、波形接続型におけるピッチ周期波形は、これを構成する基本正弦波とその多数の高調正弦波の重ねあわせととらえられるが、これら個々の振幅位相はピッチそのものに大きく依存するため、ピッチと独立した制御には適していない。

---

* “Discussion about Speech Analysis and Synthesis by Composite Wavelet Model” by Takeya SAIKACHI, Kyosuke MATSUMOTO, Shinji SAKO, and Shigeki SAGAYAMA (The University of Tokyo).

一方、CSM 合成においてはフォルマント周波数に対応する正弦波断片が、全極型フィルタにおいては単振動 (二次系) のインパルス応答である指数型減衰正弦波が、それぞれ基本波形となっており、いずれもこれら基本波形の重ねあわせと解釈できる。これら基本波形に必要な性質は、音声のスペクトル包絡をよく近似するスペクトルをもつことである。この意味からは必ずしも巡回型フィルタの場合のような長い基本波形は必要ではなく、巡回型フィルタでは単に時間特性を悪化させる要因になっていると言える。

3.2 基本波形のモデル化

以上の考察から、新しい音声分析合成方式を設計する。パラメトリックで時間特性が良い音声合成は、少なくとも有声音の合成においては、巡回型フィルタを用いず、スペクトル包絡の逆 Fourier 変換をピッチ周期で繰り返し、それに希望する振幅を乗じる方法が有効であると考えられる。さらに、合成音声の基本波形を少数の扱いやすいパラメータによって表現することができれば、音声学の知見を導入した音声加工が行なえる可能性がある。また 2.2 節の考察より、Q 値を低く抑え、かつ音声スペクトルの大きなダイナミックレンジを表現できる必要がある。

そこで、次の Fourier 変換公式に着目する。ω を周波数、t を時間、a, b, c を任意の実数とすると、

$$\mathcal{F}\left[\frac{a}{2\sqrt{b\pi}}e^{-\frac{t^2}{4b}+jct}\right] = ae^{-b(\omega-c)^2} \qquad (1)$$

が成り立つ。すなわち、周波数領域のガウス関数は、時間領域ではガウス関数と正弦波の積である Gabor 関数で表される。ガウス関数は dB 尺度で見れば下に開いた放物線であり、これを共振特性と考えると Q 値を抑えつつ、かつ大きな山と谷を形成できる点で都合がよい。

従って、音声スペクトル包絡を混合ガウス関数モデル (GMM) で近似すれば、GMM で表されたスペクトル包絡から、基本波形は Gabor 関数の重ね合わせとして容易に生成できる。各ガウス関数の平均が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ことが期待でき、分析パラメータによって音声のフォルマント構造を直接操作できる可能性がある。合成音声の品質は、スペクトル包絡の推定精度に大きく依存する。スペクトル包絡をガウス関数で近似する手法はいくつか提案されているが、本稿では予備実験の結果、亀岡らのスペクトル包絡推定法 [3] を用いた。

本手法を複合正弦波モデルの精神を継ぎ、正弦波の代わりに Gabor Wavelet の重ね合わせを基本波形とするという意味で、複合ウェーブレットモデル (Composite Wavelet Model) 法と名付ける。

4 評価実験

4.1 実験条件

提案手法の有効性を確認するために、実験によって巡回フィルタ型音声合成との時間特性の比較を行なった。実験には ATR の音声データベース B セットから女性話者の連続音声を使用した。各フレームから有声音区間を選び出して LPC フィルタで分析し、これにさまざまなピッチ周期 (原音声のピッチ周期の 0.8 倍から 1.2 倍までを 0.02 刻みで変更した) のインパルス列を 30ms のみ入力して合成を行った。そして入力停止後も合成を続け、合成音声の減衰時間[1]を調べた。CWM 法でも同様に 30ms 間の合成を行い、減衰時間を比較した。ここでは、LPC 法は 15 次元、CWM 法は 5 個のガウス関数 (15 次元) で分析合成を行った。また、両手法で分析フレーム長は 32ms、フレーム周期は 10ms とした。

[1] ここでは、入力停止から合成音声のパワー (10msec 間の振幅の二乗和) が 30dB 低下するまでの時間と定義した。

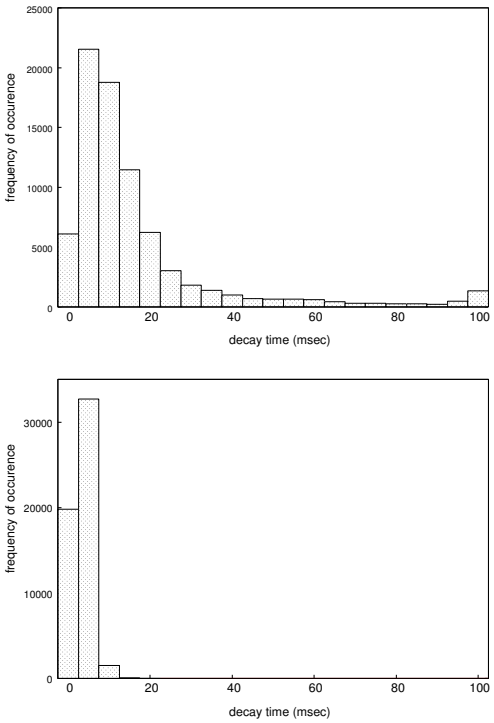

図 2: 合成波形の時間特性の傾向 (上段:LPC 法, 下段:CWM 法)

4.2 実験結果

図 2 に減衰時間を 5ms 単位のヒストグラムで示した。分布が右に偏るほど、減衰時間が長くなりやすいと言える。両者の比較により、CWM 法では明らかに減衰時間が短くなっている。

予備実験では、分析合成音の主観評価試験において、LPC 法の時間特性の問題により発生する明瞭性の低下が CWM 法により改善されていることが確認された。

5 おわりに

本稿では、従来のフィルタ型音声合成の音声品質が低下する原因として時間特性の問題があることを指摘し、これを解決するための方法である複合ウェーブレットモデルについて検討した。CWM 法では、音声のスペクトル包絡を GMM によって近似し、分析パラメータとする。そして、この GMM のフーリエ逆変換に対応する Gabor 関数を時間領域でピッチの間隔に重畳して音声を合成する。実験の結果、提案法によって合成音声の時間特性が改善することが示された。

今後は、CWM 法による合成音声の品質を改善したい。また、本研究は加工性にすぐれたテキスト音声合成を目標としており、本手法を文献 [4] の分析合成系として用いることを考えている。

参考文献

[1] 槻 武也, 松本恭輔, 酒向慎司, 嵯峨山茂樹: “ 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, vol. 105, no. 372, pp. 1-6, 2005.

[2] 嵯峨山茂樹, 板倉文忠: “複合正弦波による音声合成,” 音声研究会資料, S79-39, pp.293-300, 1979.

[3] 亀岡弘和, 小野順貴, 嵯峨山茂樹: “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, 2-6-4, 2005.

[4] 徳田恵一, 益子貴史, 小林隆夫, 今井聖: “動的特徴を用いた HMM からの音声パラメータ生成アルゴリズム,” 日本音響学会誌, vol.53, no.3, pp.192-200, 1997.