6B-01

無矛盾位相復元を用いたケプストラム特徴量からの音声合成

濱田　康弘†　　小野　順貴‡　嵯峨山　茂樹†

明治大学　総合数理学部†　　国立情報学研究所　情報学プリンシプル研究系‡

1. はじめに

テキストから音声へ変換する音声合成技術 (Text-to-Speech) は，これまでに幾つかの方法が提案されている。

波形接続型の方式では，入力テキストに従って音素や音節などの単位の波形を接続する。この方式では，実際に発声された音声波形を利用するため，自然性の高い合成音が得ることが可能だが，発声条件の変動の影響で不自然さが生じたり、接続部分に歪みが生じやすく，また，多様な声質や発声のスタイルを表現する為にはそれぞれ音声波形を必要とし，合成音声の加工性が低いという問題点がある。

そのような問題意識から，音声コーパスから得られる音響パラメータを統計的に処理する方式[1]が提案・実用化された。さらに隠れマルコフモデル(Hidden Markov model; HMM) により学習し，テキストと楽譜情報に基づいてパラメータ生成アルゴリズムを用いて合成する[2]方法が進められた。この方法では動的特徴量を考慮している為，接続部分に歪みの少ない滑らかな合成が可能であり，パラメータの変換により，多様な声質や発声のスタイルを表現することを可能とする。

しかしながら，従来の HMM 音声合成では，合成されたメルケプストラムの時間パターンから信号波形を得るために，メル対数スペクトル近似(Mel Log Spectrum Approximation; MLSA) フィルタ[3] を用いるが，巡回型フィルタであるためにインパルス応答が長くなって音声の明瞭性に影響する時間特性の問題や，基本周波数成分とスペクトル包絡のピークが重なる場合にはスペクトルのピークが鋭くなり，一部の音声振幅が不自然に大きく聴こえてしまう利得特性の問題を内在している。

一方，フィルタを用いないでパワースペクトルから信号波形を得る方法として無矛盾位相復元法[4][5] が提案されている。この方法では，パワースペクトル時系列からフレーム間で無矛盾な位相を付加することにより波形を生成するため，フィルタに起因する時間特性，利得特性の問題は生じないと考えられる。

本研究では従来の HMM から生成されたケプストラム特徴量から無矛盾位相復元を用いて合成する方法を提案する。これにより，従来の HMM 音声合成で用いられてきた巡回型フィルタで起こる時間特性・利得特性悪化の改善を試みる。

Spectral phase reconstruction applied to speech synthesis from cepstral features
†Meiji University
‡National Institute of Informatics

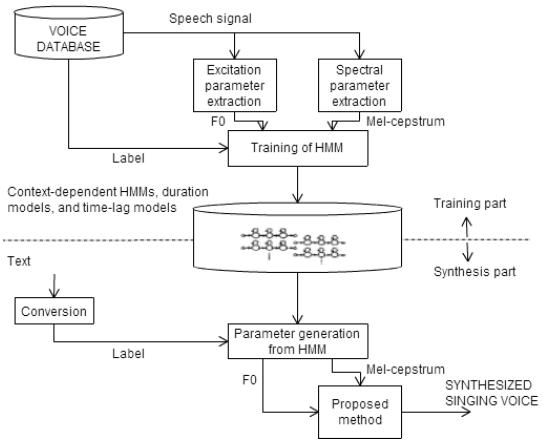

図 1．Overview of proposed HMM-based speech synthesis system.

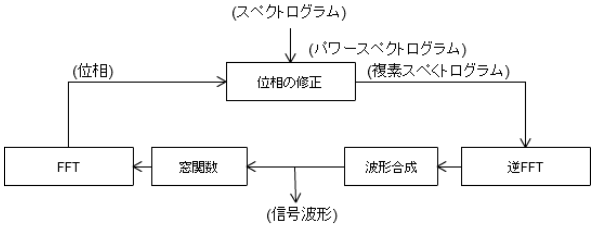

図 2．Algorithm of spectral phase reconstruction [5]

2. 非フィルタ方式の音声合成の方法

HMM 音声合成システムとして，HTS[2] で生成される一般化メルケプストラムと基本周波数を用いて算出されるパワースペクトルに無矛盾な位相を付加すれば、音声波形が得られる。この原理に基づき、以下の方法によって合成する。

(1) HTS から得られた一般化メルケプストラム係数をスペクトルに変換する。
(2) 次に，得られたスペクトル包絡から F_0 の整数倍成分のスペクトル値を抽出する。
(3) 各スペクトル値に Han 窓のスペクトルを畳み

込み，スペクトログラムを生成する。
(4) スペクトログラムからスペクトル無矛盾位相復元を用いて合成音声を生成する。

3. 実験

提案するスペクトル無矛盾位相復元による音声合成法が有効であるか調べる為に合成音声の時間特性・利得特性を調べた。比較として，MLSA フィルタによる合成音声の特性を調べた。

3.1 実験条件

実験に用いる音声は，ATR データベースより 3-5 秒程度の 5 文章を選択し，HTS により生成されたケプストラム特徴量と基本周波数からスペクトル位相復元を行った音声に対して基本周波数を 0.8 倍から 1.2 倍まで 0.05 刻みで変更したものを用いた。

3.2 時間特性の評価

有声区間 30 ms (1 フレーム)の音声を入力し，その後入力をせずに合成を行った。各音声に対して各フレーム，ピッチ周期で減衰時間を調べた。減衰時間は入力停止時から合成音声のパワーが 30 dB 低下するまでの時間とし，パワーは 10 ms 間の振幅の 2 乗和とした。

3.3 利得特性の評価

ピッチ周期を時間特性と同様に変更し，音声全体の合成を行い，有性区間の各フレームのパワーを調べた。

3.4 結果/考察

時間特性の結果を図 3 に，利得特性の結果を図 4 に示す。分布が右へ偏るほど，減衰時間が長く，利得の変化が大きい事を示している。図より，時間特性及び，利得特性は従来の巡回型フィルタに比べて改善していることが示された。

4. おわりに

本研究では，高音質な音声合成を目指し，ケプストラム特徴量からスペクトル無矛盾位相復元によって音声合成を行った。

HMM により生成されたケプストラム特徴量をスペクトルに変換し，F_0 の整数倍成分に窓関数をフーリエ変換した関数を重畳することで，パワースペクトログラムを生成し，これに対して位相復元を行った。

位相復元によって得られた合成音声の時間特性・利得特性を調べた結果，時間特性・利得特性の改善が示された。

このことから，本方法は音声合成の一手法として有効であることが示唆された。

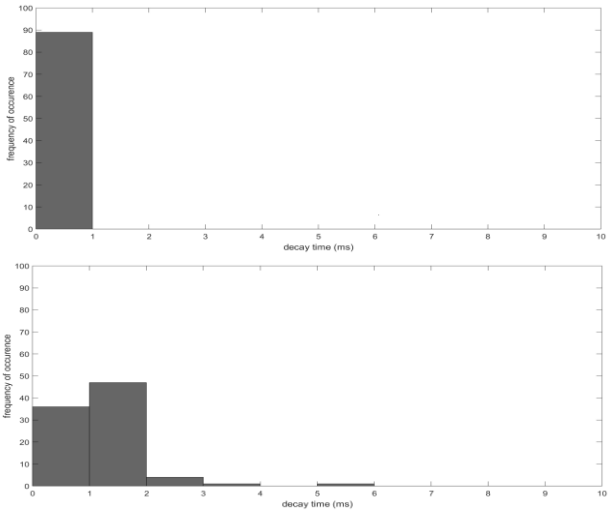

図 3. Time characteristics of proposed method (top) and MLSA filter (bottom)

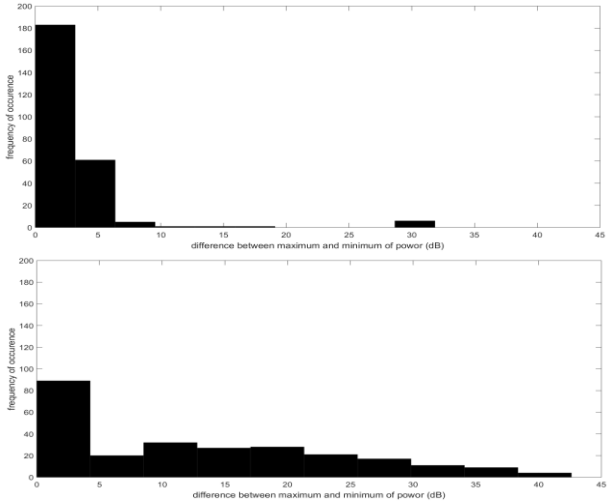

図 4. Gain characteristics of proposed method (top) and MLSA filter (bottom).

文献

[1] NTT インテリジェントテクノロジ, “高音質テキスト音声合成ボード「しゃべりん坊 HG」,” 音響誌, 49 (12), 1993.
[2] 徳田, “HMM による音声合成の基礎,” 信学論, 74, 2000.
[3] 今井他, “音声合成のためのメル対数スペクトル近似 (MLSA) フィルタ,” 信学論, J66-A (2), pp. 122-129, 1983.
[4] J. Le Roux *et al*., “Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction,” Proc. SAPA, pp. 23-28, 2008.
[5] 水野他, “パワースペクトログラムの 伸縮と無矛盾位相付加に基づく音楽音響信号の実時間テンポ／ ピッチ変換,” 音講論, pp. 843-844, 2009.