Vol.2016-MUS-111 No.39
2016/5/22

無矛盾位相復元を用いた非フィルタ方式による音声信号合成

濱田 康弘[1,a)]　小野 順貴[2,b)]　嵯峨山 茂樹[1,c)]

概要：本研究の狙いは，これまで souce-filter モデルによって行われてきた音声合成に替わる新しい音声合成法を提案することである．これまでの音声合成は von Kempelen の時代から今日に至るまで souce-filter 方式による音声合成が主流であった．パラメトリック統計的音声合成法が提案され，従来の音声合成よりも自然性の高い音声が得られるようになったが，未だ多くの課題が残されている．特に周波数領域で加工されたスペクトルを波形に戻す再合成の際に発生する品質の劣化が挙げられる．近年，パワースペクトルから無矛盾な位相を付加することによる無矛盾位相復元法が提案された．この方法を用いれば，フィルタによって発生する時間特性・利得特性悪化の問題を解決出来る可能性がある．本稿では無矛盾位相復元を用いた音声合成法を提案し，合成音声の時間特性・利得特性を調べた．その結果，従来のフィルタ方式による合成音声よりも特性が良いことが示された．このこのとから，本提案法は音声合成の一方法として有効であることが示唆された．

キーワード：音声合成，非フィルタ方式，無矛盾位相復元

1. はじめに

音声合成技術は von Kempelen の機械式音声合成器を初めとして, 長い歴史がある [1]．電気的構造のデバイスが Stewart によって発表され [2]，Dudley によって voder/vocoder が開発された [3]．その後，1960 年代後半から 1970 年代にかけて linear predictive coding (LPC)[4][5] vocoder や line spectrum pair (LSP)[6] が開発され，実用化へと繋がっていった．統計的音声合成は 1989 年に，音韻環境に基づくクラスタリングによる規則合成法 (COC 音声合成) [7][8] が開発・製品化され，その後隠れマルコフモデル (HMM) による音声合成法が確立されていった [9].

これまでの音声合成法の多くは人の発話機構を模擬する source-filter モデルが用いられている．その理由の一つとして，von Kempelen から続く人の発話機構に基づく音声合成がなされてきているという歴史的背景があるが，もう一つの理由として，パラメトリック音声合成では，波形に戻す際に source-filter に基づく再合成手法を用いる必要性がある為である．

[1] 明治大学
Meiji University, 4-21-1, Nakano, Tokyo, Japan
[2] 国立情報学研究所 / 総合研究大学院大学
National Institute of Informatics / Graduate University for Advanced Studies
[a)] hamada@meiji.ac.jp
[b)] onono@nii.ac.jp
[c)] sagayama@meiji.ac.jp

近年，パワースペクトルから波形を生成する無矛盾位相復元法が開発された [10][11]．これは，隣接するフレーム間でのパワースペクトルと位相スペクトルに整合性を保つ無矛盾な位相を反復演算により推定し，波形を生成する方法である．

本研究の狙いは，これまで source-filter モデルによって行われてきた音声合成に替わる新しい音声合成法を提案することである．source-filter による音声合成の品質劣化の一要因は巡回型フィルタであることによって引き起こされる時間特性と利得特性の悪化が挙げられる．無矛盾位相復元を用いれば，非フィルタ方式である為，フィルタによって引き起こされる問題を解消できる可能性がある．

2. 非フィルタ方式による音声合成

2.1 非フィルタ方式による音声合成法の概要

本稿では HMM 音声合成システムをベースとし，合成部に焦点をあてる．典型的な HMM 音声合成システムではテキスト情報を入力とし，HMM によりケプストラム特徴量と F_0 特徴量が生成される (2.2 節)．ケプストラム特徴量と F_0 特徴量から所望のパワースペクトルを算出出来れば，無矛盾位相復元によって波形生成が可能である (2.5 節)．スペクトル包絡はケプストラム特徴量からスペクトルに変換することで得ることが出来る (2.3 節)．有声音の声帯振動の基本周期は周波数領域で F_0 の整数倍の調波構造として表れる為，スペクトル包絡を F_0 の整数倍成分でサンプリ

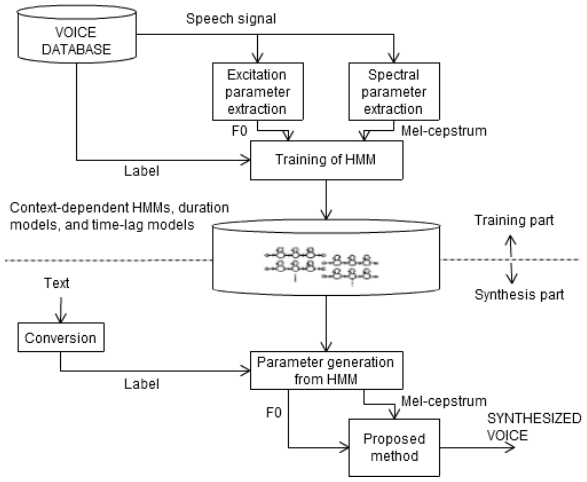

図 1 Overview of proposed HMM-based synthesis system

ングしたものに相当し，時間領域での掛け算は周波数領域での畳み込みに対応することから，所望のスペクトルはスペクトル包絡を F_0 の整数倍成分でサンプリングした線スペクトルと周波数領域での窓関数との畳み込みの近似として得ることが出来る (2.4 節).

2.2 HMM を基にしたケプストラム特徴量と F_0 の生成

提案手法の特徴パラメータ学習・生成は典型的な HMM 音声合成システムに従って行われた [12][13]．合成部において，典型的な HMM 音声合成システムではメル対数スペクトル近似 (MLSA) フィルタ [14] を用いているのに対し，提案手法では無矛盾位相復元が用いられた (図 1)．HMM 音声合成システムではスペクトルパラメータとしてメルケプストラムが用いられ，励起パラメータとして F_0 が用いられる．合成部では与えれたテキスト情報からこれらのパラメータが推定された．

2.3 ケプストラム特徴量からスペクトルへの変換

生成されたメルケプストラムは次のようにスペクトルに変形される．

$$H(\omega) = s_\gamma^{-1}\left(\sum_{m=0}^{M} \tilde{c_\gamma}(m) e^{-j\tilde{\omega}t}\right) \tag{1}$$

ここで，

$$s_\gamma^{-1}(\omega) = \begin{cases} (1+\gamma\omega)^{1/\gamma}, & 0 < |\gamma| \leq 1 \\ \exp\omega, & \gamma = 0 \end{cases} \tag{2}$$

$$\tilde{\omega} = \tan^{-1}\frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha} \tag{3}$$

$H(\omega)$ はスペクトル，s_γ^{-1} は一般化対数関数の逆関数，$\tilde{c_\gamma}(m)$ はメル一般化ケプストラムである．α は周波数圧縮パラメータであり，メル一般化ケプストラムは冪パラメータ $\gamma = 0$ でケプストラム，$\gamma = -1$ で AR 係数に対応する．

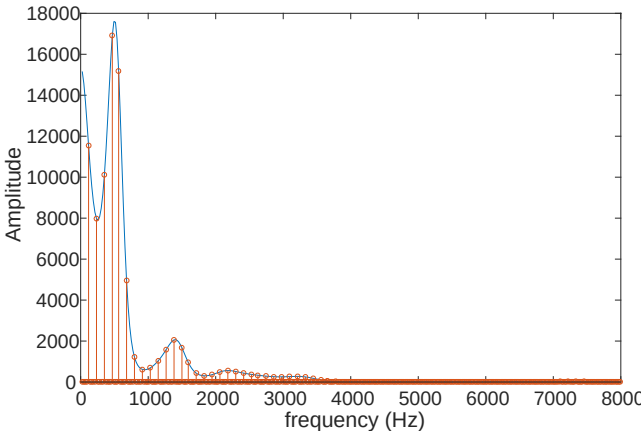

図 2 Harmonic components of a spectrum

2.4 基本周波数成分の記述

HMM より生成された F_0 を用いて，$H(\omega)$ から F_0 の整数倍成分のスペクトル値を抽出し，スペクトル $H(\omega)$ に微細構造を与える．図 2 に F_0 が 120 Hz である時のスペクトルのサンプルを示す．

F_0 の調波成分を含んだスペクトルは次のように $H(\omega)$ から F_0 の整数倍成分が抽出されたスペクトル値と Hann 窓をフーリエ変換したスペクトル $W(\omega)$ の畳み込みによって得られる．Hann 窓のスペクトルは図 3 のようにサイドローブの減衰が鋭いため，より調波構造を再現できると考え適用した．

$$w(n) = 0.5(1 - \cos(\frac{2\pi n}{N-1})), n = 1, 2, ...N \tag{4}$$

$$W(\omega) = \exp(j\omega\frac{N-1}{2})\frac{\sin(N\omega/2)}{\sin(\omega/2)} \tag{5}$$

$$X(\omega) = H(\omega) * W(\omega) \tag{6}$$

図 3 に Hann 窓 (上) と Hann 窓のスペクトル (下) を，図 4 に $H(\omega)$ と $W(\omega)$ が畳み込まれたスペクトログラムを示す．

2.5 スペクトル無矛盾位相復元

合成音声波形は，得られたパワースペクトルから無矛盾位相復元を用いて生成される．図 5 に示すように，無矛盾位相復元では短時間フーリエ変換と逆短時間フーリエ変換を反復処理し，位相を更新することで無矛盾な位相を有する波形が生成される．

3. 時間・利得特性の評価

3.1 時間特性と利得特性の問題

通常，分析合成系において再合成された音声の自然性は元の音声よりも自然性が低くなる．その品質の劣化の一要因として，合成音の時間特性と利得特性が関係していると考えられる．MLSA フィルタのような巡回型フィルタを用いると，フォルマントの Q 値が F_0 の調波構造と重なるこ

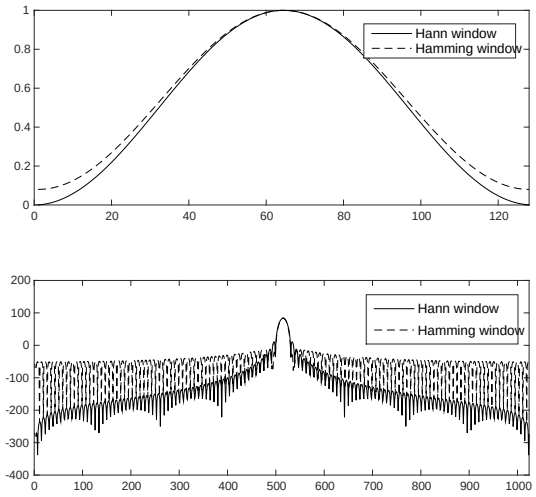

図 3 Hann window (top) and spectrum of Hann window (bottom)

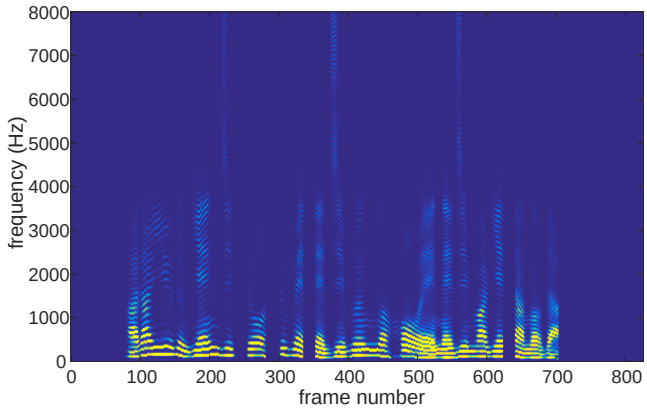

図 4 Spectrogram of convolved $H(\omega)$ with $W(\omega)$

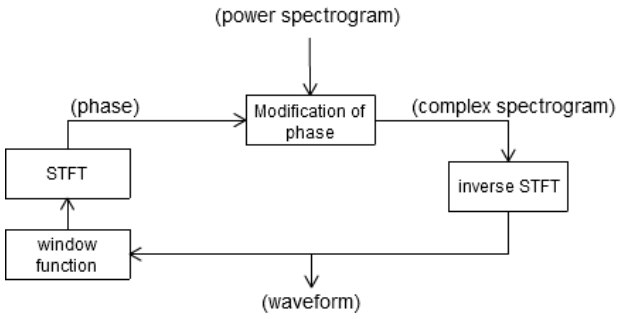

図 5 Algorithm of spectral phase reconstruction

とにより，出力の振幅が減衰するのに時間がかかってしまう．また，出力の利得は Q 値に比例するので，利得が F_0 によって変動してしまう．提案手法はフィルタを用いない手法であるので，これらの時間特性・利得特性の問題を解消できる可能性がある．そこで，提案手法と MLSA フィルタを用いた手法の時間特性，利得特性を調べた．

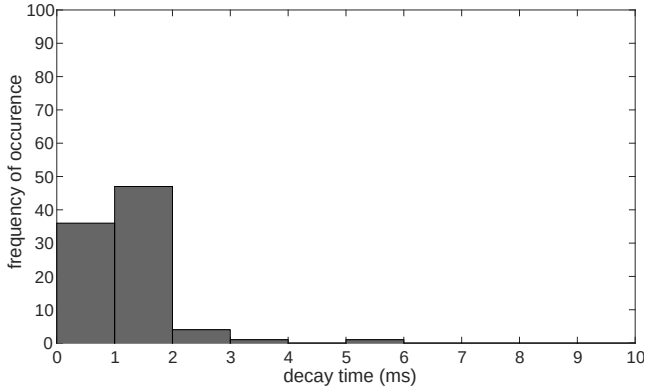

図 6 Time characteristics of MLSA filter

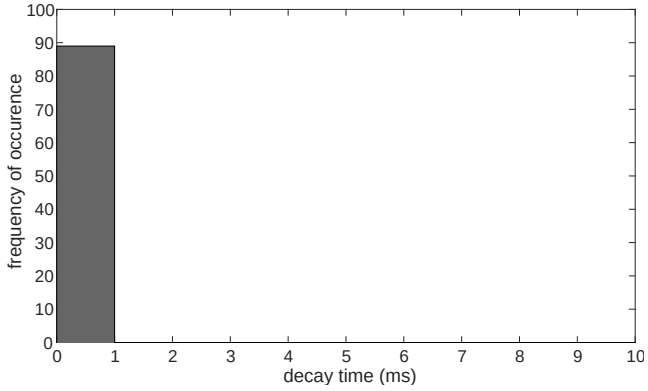

図 7 Time characteristics of proposed method

3.2 実験条件

実験に用いた音声は，5 文章をテキスト情報として入力し，HTS[13] を用いて生成されたケプストラム特徴量と基本周波数から合成された音声であった．ケプストラム特徴量は，$\gamma = 1.0$ で，$\alpha = 0.55$，サンプリング周波数は 48000 Hz であり，フレームシフトは 5 ms で，フレーム長は 40 ms で分析された．合成音声の継続時間長は 3.35 から 5.15 s であり，基本周波数は，各手法において 0.8 倍から 1.2 倍まで 0.05 刻みで変更された．

3.3 時間特性の評価/結果

有声区間 30 ms (1 フレーム) の音声を入力し，その後入力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 間の振幅の二乗和とした．

時間特性の結果を図 6，7 にヒストグラムで示す．図は分布が右へ偏るほど，減衰時間が長い事を示している．図より，時間特性は MLSA フィルタに比べて提案手法の方が良いことが示された．

3.4 利得特性の評価/結果

基本周波数を時間特性と同様に変更し，音声全体の合成を行い，有声区間の各フレームのパワーを調べた．

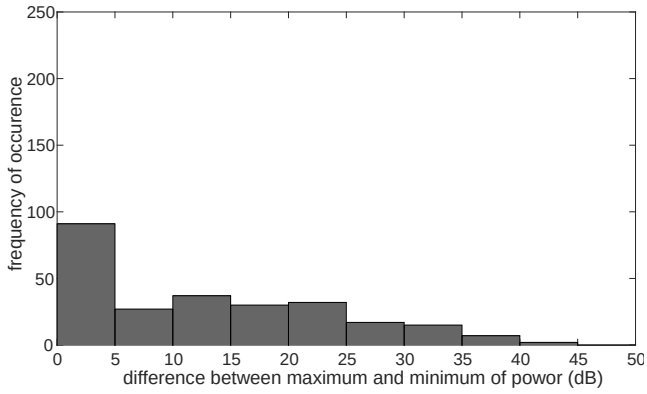

図 8 Gain characteristics of MLSA filter

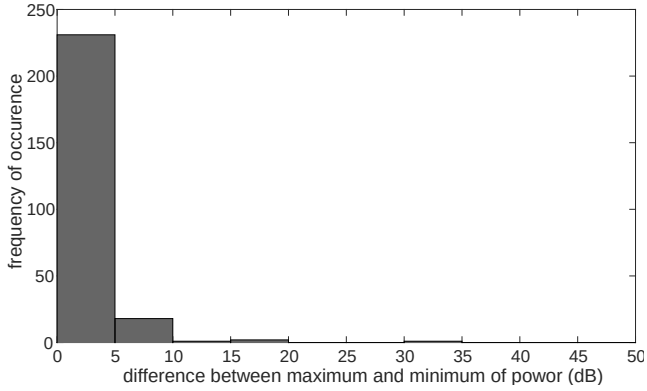

図 9 Gain characteristics of proposed method

利得特性の結果を図 8，9 にヒストグラムで示す．図は分布が右へ偏るほど，利得の変化が大きい事を示している．図より，利得特性は MLSA フィルタに比べて提案手法の方が良いことが示された．

3.5 考察

提案手法はフィルタを用いていないため，時間特性・利得特性の悪化する要因がなく，巡回型フィルタである MLSA フィルタでは時間特性・利得特性が悪化している．これは，合成音声のフォルマントの Q 値が F_0 の整数倍成分の重なることにより，減衰時間が長く，利得特性の変動が異常に大きくなっている為だと考えられる．予備的聴取実験を行ったところ，実験参加者の聴取印象から，MLSA フィルタを用いた合成音声に比べ提案手法の合成音声は歯切れの良い，清涼感のあるという意見があり，時間特性・利得特性の影響が合成音声の自然性に反映されたものであると考えられる．

4. おわりに

本稿では，source-filter 方式の音声合成方式に替わる無矛盾位相復元を用いた非フィルタ方式による音声合成の新しい方法を提唱した．HMM によって学習・生成されたケプストラム特徴量と基本周波数からパワースペクトルを生成し，無矛盾位相復元による波形生成を行った．時間特性・利得特性を調べた結果，フィルタを用いた合成法より提案した合成法の特性が良い結果が得られた．このことから，提案した非フィルタ方式による方法は音声合成の一手法としての有効性が示唆された．今後，定量的な自然性の主観評価を行い，合成音の品質を調べるとともに，パラメトリック統計音声合成で問題となっている平均化問題に対して，学習する特徴量を検討していく予定である．

謝辞 この研究は、日本学術振興会科学研究費補助金基盤研究 (A) 課題番号 26240025「音楽信号・作曲理論・演奏の数理モデルを融合する音楽音響情報処理の研究」の部分的支援を得て行われた．

参考文献

[1] von Kempelen, W.: *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine* (1791).
[2] John.Q.Stewart: An Electrical Analogue of the Vocal Organs, *Nature*, Vol. 110, No. 7, pp. 311–312 (1922).
[3] Dudley, H.: Remaking Speech, *The Journal of neuroscience : the official journal of the Society for Neuroscience*, Vol. 11, No. 2, pp. 169–177 (1939).
[4] Itakura, F. and Saito, S.: Analysis synthesis telephony based on the maximum likelihood method, *Reports of the 6th Int.Cong.Acoust*, No. C-5-5 (1968).
[5] Atal, B. and Schroeder, M.: Predictive coding of speech signals, *Reports of the 6th Int.Cong.Acoust*, No. C-5-4 (1968).
[6] Sagayama, S.: Stability condition of LSP speech synthesis digital filters, *Proceedings Acout.Soc.Japan*, No. 3-4-12 (1972).
[7] 中嶌信弥，浜田洋：音韻環境に基づくクラスタリングによる規則合成法，電子情報通信学会論文誌，Vol. J72-D-II, No. 8, pp. 1174–1179 (1989).
[8] (株)NTT インテリジェントテクノロジ：高品質テキスト音声合成ボード「しゃべりん坊 HG」，日本音響学会誌，Vol. 49, No. 12, p. 881 (1993).
[9] 徳田恵一，小林隆夫，千葉健司，今井聖：メル一般化ケプストラム分析による音声のスペクトル推定，電子通信学会論文誌，Vol. 75, No. 7, pp. 1124–1134 (1992).
[10] Griffin, D.: Signal estimation from modified short-time Fourier transform, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 32, No. 2, pp. 236–243 (1984).
[11] Roux, J. L., Ono, N. and Sagayama, S.: Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction, *Proc. SAPA*, No. Sapa, pp. 23–28 (2008).
[12] Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J. and Oura, K.: Speech synthesis based on hidden Markov models, *Proceedings of the IEEE*, Vol. 101, No. 5, pp. 1234–1252 (2013).
[13] Zen, H., Nose, T., Yamagishi, J. and Sako, S.: The HMM-based speech synthesis system (HTS) version 2.0., *Proc. SSW6*, pp. 294–299 (2007).
[14] 今井聖，住田一男，古市千枝子：音声合成のためのメル対数スペクトル近似 (MLSA) フィルタ，信号処理学会論文誌，Vol. J66-A, No. 2, pp. 122–129 (1983).