変調スペクトルを考慮したHMM音声合成*

◎ 高道 慎之介, 戸田 智基, Graham Neubig, Sakriani Sakti, 中村 哲 (奈良先端大)

1 はじめに

HMM音声合成 [1] において，生成されるパラメータ系列の過剰な平滑化は，音質劣化の一因となる．これに対して本稿では，パラメータ系列の変調スペクトル （MS: Modulation Spectrum）に基づくポストフィルタを提案する．提案法では，生成パラメータ系列のMSが，自然音声のパラメータ系列のMSと近くなるように，生成パラメータ系列に対してフィルタ処理を施す．実験的評価によりその有効性を示す．

2 HMM音声合成のパラメータ生成法

2.1 HMM尤度最大化基準 [2]

HMM音声合成では，自然音声のパラメータ系列からコンテキスト依存HMMを学習する．生成時には，合成対象のテキストに対応する文HMMを形成し，静的・動的特徴量間の明示的な制約条件の下でHMM尤度を最大化することで，パラメータ系列を生成する．

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmax}}\, P\left(\boldsymbol{W}\boldsymbol{c}|\boldsymbol{\lambda}\right) \tag{1}$$

ただし，$\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \cdots, \boldsymbol{c}_T^\top\right]^\top$ はTフレームの音声パラメータ系列，$\boldsymbol{c}_t = \left[c_t\left(1\right), \cdots, c_t\left(D\right)\right]^\top$ は時刻tにおけるD次元の音声パラメータ，$\boldsymbol{W}$は動的特徴量の計算に用いる重み係数によって決定される行列 [2]，$\boldsymbol{\lambda}$はHMMのパラメータセットを表す．

式 (1) により生成されるパラメータは，過剰に平滑化される傾向にある．

2.2 HMM尤度・GV尤度最大化基準 [3]

系列内変動（GV: Global Variance）は，パラメータ系列全体の変動成分を表し，次式で定義される．

$$\boldsymbol{v}\left(\boldsymbol{c}\right) = \left[v\left(1\right), \cdots, v\left(D\right)\right]^\top \tag{2}$$

$$v\left(d\right) = \frac{1}{T}\sum_{t=1}^{T}\left(c_t\left(d\right) - \frac{1}{T}\sum_{\tau=1}^{T} c_\tau\left(d\right)\right)^2 \tag{3}$$

生成時には，HMM尤度及びGV尤度を最大化してパラメータ系列を生成する．

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmax}}\, P\left(\boldsymbol{W}\boldsymbol{c}|\boldsymbol{\lambda}\right) P\left(\boldsymbol{v}\left(\boldsymbol{c}\right)|\boldsymbol{\lambda}_v\right)^w \tag{4}$$

ただし，$\boldsymbol{\lambda}_v$はGVの確率密度関数のパラメータセット，wはGV尤度の重みを表す．

生成パラメータ系列のGVは，GV尤度を考慮することで補償され，合成音声の音質が改善する．

3 変調スペクトルを考慮したHMM音声合成

3.1 パラメータ系列の変調スペクトル分析

MSは，本来，パラメータ系列をフーリエ変換した値を表す [4] が，本稿では，その対数振幅スペクトルをMSと呼ぶ．パラメータ系列$\boldsymbol{c}$に対するMSを次式で定義する．

$$\boldsymbol{s}\left(\boldsymbol{c}\right) = \left[\boldsymbol{s}\left(1\right)^\top, \cdots, \boldsymbol{s}\left(D\right)^\top\right]^\top \tag{5}$$

$$\boldsymbol{s}\left(d\right) = \left[s_d\left(0\right), \cdots, s_d\left(M\right)\right]^\top \tag{6}$$

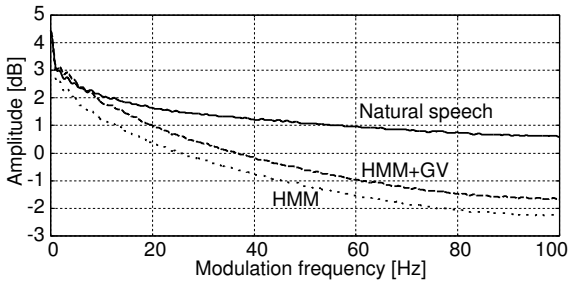

Fig. 1 メルケプストラム系列の変調スペクトル

ただし，$s_d\left(m\right)$は，d次元目のパラメータ系列$\left[c_1\left(d\right), \cdots, c_T\left(d\right)\right]^\top$に対する，周波数インデックス$m$のMS，$M$は離散フーリエ変換のサンプル数の半分を表す．本稿では発話毎にMSを計算する．

Fig. 1に，式 (1)（HMM）と式 (4)（HMM+GV）で生成された第10次メルケプストラム系列のMSを示す．比較のため，自然音声（Natural speech）の同系列のMSも示す．“HMM” のMSは，自然音声のパラメータ系列のMSと比較して，大きく減衰していることが確認できる．また，“HMM+GV” のMSは，GVの導入により比較的補償されるものの，未だに大きく減衰している．故に，MSの直接的な補償により，合成音声の音質改善がもたらされると期待される．

3.2 変調スペクトルに基づくポストフィルタ

MSを補償するポストフィルタ処理を提案する．ポストフィルタは，学習データを用いて事前に設計する．

3.2.1 学習部

自然音声のパラメータ系列から，次式に示す確率密度関数を学習する．

$$P\left(\boldsymbol{s}\left(\boldsymbol{c}\right)|\boldsymbol{\lambda}_s\right) = \mathcal{N}\left(\boldsymbol{s}\left(\boldsymbol{c}\right); \boldsymbol{\mu}^{(\mathrm{N})}, \boldsymbol{\Sigma}^{(\mathrm{N})}\right) \tag{7}$$

ただし，$\mathcal{N}\left(\cdot; \boldsymbol{\mu}^{(\mathrm{N})}, \boldsymbol{\Sigma}^{(\mathrm{N})}\right)$は平均$\boldsymbol{\mu}^{(\mathrm{N})} = \left[\mu_{1,0}^{(\mathrm{N})}, \cdots, \mu_{D,M}^{(\mathrm{N})}\right]^\top$と対角共分散行列$\boldsymbol{\Sigma}^{(\mathrm{N})} = \mathrm{diag}\left[\left(\sigma_{1,0}^{(\mathrm{N})}\right)^2, \cdots, \left(\sigma_{D,M}^{(\mathrm{N})}\right)^2\right]$の正規分布，$\mu_{d,m}^{(\mathrm{N})}$と$\left(\sigma_{d,m}^{(\mathrm{N})}\right)^2$は$s_d\left(m\right)$の平均と分散，$\boldsymbol{\lambda}_s$はMSの確率密度関数のパラメータセットを表す．同様に，HMM音声合成で生成されたパラメータ系列から正規分布$\mathcal{N}\left(\cdot; \boldsymbol{\mu}^{(\mathrm{G})}, \boldsymbol{\Sigma}^{(\mathrm{G})}\right)$を学習する．なお，自然音声のパラメータ系列と生成パラメータ系列間の継続長の違いがMSに影響することを避けるために，正規分布の学習に用いる生成パラメータ系列は，自然音声の継続長において生成する．

3.2.2 生成部

生成されたパラメータ系列$\boldsymbol{c}$のMSに対して次式のポストフィルタを適用する．

$$s'_d\left(m\right) = \left(1-k\right)s_d\left(m\right) + k\left[\frac{\sigma_{d,m}^{(\mathrm{N})}}{\sigma_{d,m}^{(\mathrm{G})}}\left(s_d\left(m\right) - \mu_{d,m}^{(\mathrm{G})}\right) + \mu_{d,m}^{(\mathrm{N})}\right] \tag{8}$$

*HMM-based speech synthesis considering modulation spectrum, by TAKAMICHI, Shinnosuke, TODA, Tomoki, SAKTI, Sakriani, NEUBIG, Graham, NAKAMURA, Satoshi (NAIST)

ただし，k はポストフィルタ強度係数 $(0 \leq k \leq 1)$ を表す．フィルタ後の MS は，$k = 1$ の際には自然音声のパラメータ系列の MS に近い値となり，$k = 0$ の際にはポストフィルタ処理前と等価となる．ポストフィルタ後のパラメータ系列は，式 (8) で計算された MS と，フィルタ処理前のパラメータ系列の周波数位相特性から計算する．

4 実験的評価

4.1 実験条件

学習データは女性話者による ATR 音素バランス文 [5] A-I セット 450 文とする．評価データは同 J セット 53 文を使用する．学習データのサンプリング周波数は 16 kHz，フレームシフトは 5 ms とする．スペクトル特徴量は，STRAIGHT 分析 [6] による 0 次から 24 次のメルケプストラム係数，音源特徴量は，対数 F_0，5 周波数帯域における平均非周期成分を使用する．5 状態 left-to-right 型の隠れセミマルコフモデル（Hidden Semi-Markov Model : HSMM）の学習を行う．変調スペクトルにおける離散フーリエ変換のサンプル数は 4096 点とする．これは，学習・評価データのパラメータ系列のフレーム数を十分に超える値である．提案法はスペクトルパラメータに対してのみ適用し，音源パラメータは式 (1) で生成する．

以下に示す手法を用いて評価を行う．

HMM： 式 (1) で生成

HMM+MS： 式 (1) で生成したパラメータ系列に対して提案法を適用

HMM+GV： 式 (4) で生成

HMM+GV+MS： 式 (4) で生成したパラメータ系列に対して提案法を適用

まず，ポストフィルタ強度係数を決定するための評価を行う．ポストフィルタ強度係数を 0 から 1 まで 0.05 刻みで変化させ，ポストフィルタ処理後のパラメータ系列に対する HMM 尤度，GV 尤度及び MS 尤度を計算する．同時に，自然音声（Natural speech）のパラメータ系列に対する尤度も計算する．

次に，提案法による音質改善効果を対比較実験により評価する．評価は 8 人の受聴者に対するプリファレンススコアとする．

4.2 客観評価結果

ポストフィルタ強度係数を変化させた時の，ポストフィルタ後のパラメータ系列に対する HMM 対数尤度を Fig. 2 に，GV 対数尤度を Fig. 3 に，MS 対数尤度を Fig. 4 に示す．Fig. 2 から，ポストフィルタ強度係数を大きくするに従い，生成パラメータ系列に対する HMM 尤度は大きく減少することがわかる．しかしながら，その尤度は自然音声のパラメータ系列に対する HMM 尤度よりも依然として大きい．一方，Fig. 3 から，ポストフィルタ強度係数を大きくするに従い GV 尤度は変化し，ポストフィルタ強度係数を 0.85 に設定した場合に，“HMM+MS” と “HMM+GV+MS” の両方の尤度が自然音声に接近していることがわかる．対して Fig. 4 から，生成パラメータ系列に対する MS 尤度は，自然音声のパラメータ系列に対する尤度よりも常に小さいことがわかる．以上の結果から，ポストフィルタ強度係数を 0.85 に設定する．

4.3 主観評価結果

音質の主観評価結果を Fig. 5 に示す．“HMM” における生成パラメータ系列に対して提案法を適用することで，スコアが著しく上昇し，“HMM+GV” と同等の音質が得られることが分かる．また，“HMM+GV” におけるパラメータ系列に対する提案法の適用により，スコアは更に上昇することがわかる．以上の結果から，提案法による音質の改善が確認できる．

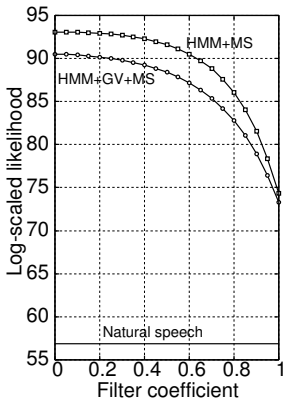

Fig. 2 HMM 尤度

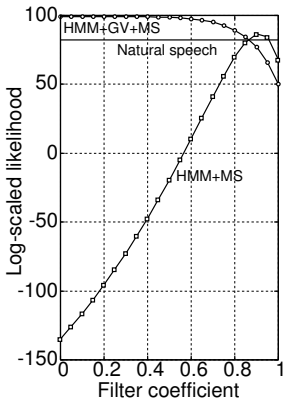

Fig. 3 GV 尤度

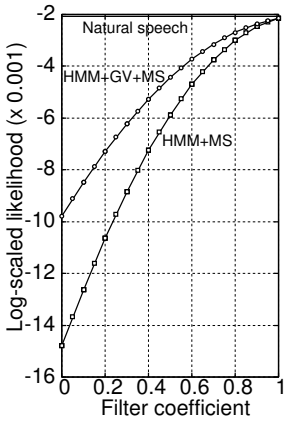

Fig. 4 MS 尤度

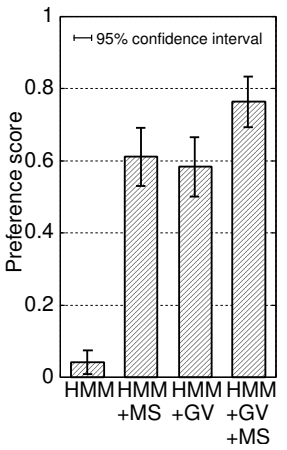

Fig. 5 主観評価結果

5 まとめ

本稿では，HMM 音声合成の音質改善を目的として，従来の HMM 音声合成の生成パラメータ系列に対して，変調スペクトルを補償するポストフィルタ処理を提案し，その有効性を実験的評価により示した．今後は，変調スペクトルを考慮したパラメータ生成法の検討を行う．

謝辞 本研究の一部は，JSPS 科研費 22680016 の助成を受け実施した．また，本研究の一部は，(独）情報通信研究機構の委託研究「知識・言語グリッドに基づくアジア医療交流支援システムの研究開発」の一環として実施した．

参考文献

[1] H. Zen *et al.*, *Speech Commun.*, 51(11), pp. 1039–1064, 2009.
[2] K. Tokuda *et al.*, Proc. *ICASSP*, Vol. 3, pp. 1315–1318, 2000.
[3] T. Toda *et al.*, *IEICE Transactions*, Vol. E90–D, No. 5, pp. 816–824, 2007.
[4] L. Atlas *et al.*, *EURASIP Journal*, Vol. 7, pp. 668–675, 2003.
[5] 阿部 他, ATR テクニカルレポート, TR–I–0166, 1990.
[6] H. Kawahara *et al.*, Speech Communication, Vol. 27, No. 3–4, pp. 187–207, 1999.