Vol.2015-MUS-107 No.71
2015/5/24

統計的パラメトリック音声合成のための変調スペクトルに基づく音質改善法

高道 慎之介[1,2,a)]　戸田 智基[1]　ブラック アラン[2]　中村 哲[1]

概要：統計的パラメトリック音声合成方式は，合成器構築の容易さ及びその汎用性の高さから，話声のみならず歌声の合成・変換法として広く利用されている．しかしながら，その合成音声の音質は，自然音声の音質と比較すると未だに大きく劣化する傾向にある．本稿では，合成音声の音質改善を目的として我々が提案している，変調スペクトルに基づく３つの方法について紹介する．

キーワード：統計的パラメトリック音声合成，HMM 音声合成，GMM 声質変換，過剰な平滑化，変調スペクトル

Quality Improvement Approaches Based on the Modulation Spectrum to Statistical Parametric Speech Synthesis

Shinnosuke Takamichi[1,2,a)]　Tomoki Toda[1]　Alan W. Black[2]　Satoshi Nakamura[1]

Abstract: Statistical parametric speech synthesis allows us to produce speaking and singing voice by utilizing easy building of the speech synthesizer and the high versatility. However, the synthetic speech quality is significantly degraded compared to natural speech. This paper introduces 3 approaches based on the modulation spectrum for high-quality statistical parametric speech synthesis.

Keywords: statistical parametric speech synthesis, HMM-based speech synthesis, GMM-based voice conversion, over-smoothing, modulation spectrum

1. はじめに

入力情報から音声を生成する音声合成技術は，我々のコミュニケーション能力や身体機能を拡張する可能性を秘めている．本稿で取り扱う，テキストから音声を合成するテキスト音声合成 [1] と言語情報を保持しつつ声質を変換する声質変換 [2] は，音声合成技術の代表例である．1990 年代に提案され 2000 年代に急速に普及した統計的パラメトリック音声合成方式 [3], [4], [5] は，合成器構築の容易さ及びその汎用性の高さ [6], [7] から，身障補助 [8], [9]，言語教育支援 [10], [11] 等に向けて広く研究されている．更に近年，話声のみならず，歌声の合成・変換を可能にする枠組み [12], [13] が研究されるなど，この合成方式の応用範囲の枚挙に遑がない．

一方で，統計的パラメトリック音声合成方式における合成音声の音質は，自然音声と比較して著しく劣化する傾向にある [14], [15]．その要因は，分析部・学習部・生成部の各々に存在する [16] が，特に生成部では，統計処理による音声パラメータ系列の過剰な平滑化が大きな要因である．近年，我々は，音声パラメータ系列の変調スペクトル（MS: Modulation Spectrum）が，過剰な平滑化の定量化に効果的であることを明らかにしている [17]．MS は，音声パラメータ系列のパワースペクトルとして定義され，音声知覚 [18] ・音声認識 [19] ・音声識別 [20] にも利用される特徴量である．本稿では，HMM（Hidden Markov Model）音声合成 [3] と GMM（Gaussian Mixture Model）声質変換

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
[2] カーネギーメロン大学
Carnegie Mellon University
[a)] shinnosuke-t@is.naist.jp

[4] において，MS に基づく 3 つの音質改善法を紹介する．

2. 統計的パラメトリック音声合成

HMM 音声合成と GMM 声質変換の学習部及び生成部について概説する．

2.1 学習部

入力情報の特徴量系列 $\boldsymbol{X}$（HMM 音声合成における入力テキストのコンテキスト又は GMM 声質変換における入力音声の特徴量）と出力音声の特徴量系列 $\boldsymbol{Y}$ から，統計モデルパラメータ $\boldsymbol{\lambda}$ を次式の様に最尤推定する．

$$\boldsymbol{\lambda} = \begin{cases} \operatorname{argmax} P\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\lambda}\right) & \text{(HMM)} \\ \operatorname{argmax} P\left(\boldsymbol{Y}_t,\boldsymbol{X}_t|\boldsymbol{\lambda}\right) & \text{(GMM)} \end{cases} \tag{1}$$

ただし，$\boldsymbol{X}_t$ と $\boldsymbol{Y}_t$ はそれぞれ，時刻 t における入出力音声の特徴量である．$\boldsymbol{Y}$ は，静的特徴量とその時間変化を表す動的特徴量から成り，$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}$ で表現される [16]．ただし，$\boldsymbol{y}$ は音声パラメータ系列，$\boldsymbol{W}$ は，$\boldsymbol{y}$ から静的・動的特徴量を計算する行列である．

2.2 生成部

生成時には，入力コンテキスト系列又は入力音声パラメータ系列 $\boldsymbol{X}$ から，対応する HMM 又は GMM 系列を構築し，次式に示すように，条件付き確率を最大にするように音声パラメータ系列 $\hat{\boldsymbol{y}}$ を生成する．

$$\hat{\boldsymbol{y}} = \operatorname{argmax} P\left(\boldsymbol{W}\boldsymbol{y}|\boldsymbol{X},\hat{\boldsymbol{q}},\boldsymbol{\lambda}\right) \tag{2}$$

ただし，$\hat{\boldsymbol{q}}$ は，HMM 音声合成における準最適な HMM 状態系列又は GMM 声質変換における GMM 分布系列であり，HMM 状態継続長モデルの尤度最大化 [21] 又は周辺化 GMM の事後確率最大化 [4] により決定される．この式で生成されるパラメータ系列は，統計処理による過剰な平滑化の影響を強く受けるが，系列内変動（GV: Global Variance）[4], [22] を考慮することでその影響を比較的緩和できる．

3. 変調スペクトルに基づく音質改善法

MS に基づく 3 つの音質改善法を紹介する．各手法の適用箇所は図 1 参照とする．

3.1 変調スペクトル (MS)

MS は，GV の拡張であり，パラメータ系列のパワースペクトルとして定義される [17]．具体的には，音声パラメータの各次元の系列をフーリエ変換し，その二乗を計算した特徴量である．本稿では，パラメータ系列 $\boldsymbol{y}$ の MS を $\boldsymbol{s}\left(\boldsymbol{y}\right)$ で表す．生成パラメータ系列は統計処理により時間方向に平滑化されるため，その MS は自然音声パラメータの MS と比較して減衰する．以降に示す 3 手法は，生成パラメータ系列の MS を補償することで合成音声の音質を改善する．

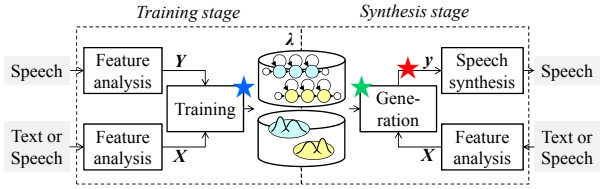


図 1 統計的パラメトリック音声合成の処理手順と，3 つの音質改善法の適用箇所（赤，緑，青星印はそれぞれ，手法 1, 2, 3 の適用箇所を表す．）

Fig. 1 Procedure of statistical parametric speech synthesis and the point that applied the MS. The red, green, and blue star indicate the proposed methods 1, 2, and 3, respectively.

3.2 手法 1： MS を補償するポストフィルタ [17], [23], [24]

生成パラメータ系列の MS を補償する最も簡易な方法は，生成パラメータ系列の直接的な変形である．提案するポストフィルタは，MS の線形変換処理により生成パラメータ系列の MS を補償する．ポストフィルタのパラメータは，音声合成に用いる学習データから自動的に推定される．統計処理による MS の減衰は，スペクトルだけでなく F_0 と音素継続長においても観測されるため，多様な音声特徴量に対するポストフィルタ処理が可能である[*1]．また，この処理は，音声合成器の学習・生成処理から独立した処理であるため，多様な音声合成方式に対して適用可能である．

3.3 手法 2: MS を考慮した音声パラメータ生成 [26], [27]

手法 1 のフィルタ処理は，従来の生成基準を無視した変形を行うため，生成パラメータ系列を過剰に変形し過強調された音声を生成する．そこで，次式に示すように，従来の生成基準と MS 基準を同時に最適化して，音声パラメータ系列を生成する．

$$\hat{\boldsymbol{y}} = \operatorname{argmax} P\left(\boldsymbol{W}\boldsymbol{y}|\boldsymbol{X},\hat{\boldsymbol{q}},\boldsymbol{\lambda}\right) P\left(\boldsymbol{s}\left(\boldsymbol{y}\right)|\boldsymbol{\lambda}_{\mathrm{s}}\right)^{\omega_{\mathrm{s}}} \tag{3}$$

ただし，$\boldsymbol{\lambda}_{\mathrm{s}}$ と ω_{s} はそれぞれ，MS のモデルパラメータセットと MS 尤度の重みを表す．MS の確率密度関数 $P\left(\boldsymbol{s}\left(\boldsymbol{y}\right)|\boldsymbol{\lambda}_{\mathrm{s}}\right)$ は，学習データを用いて推定される．この枠組みにより，従来の生成基準を満たした MS 補償が可能である．図 2 に生成パラメータ系列の MS の例を示す．HMM 音声合成における従来の生成法（“HMM”）[21] と GV を考慮した生成法（“HMM+GV”）[22] と比較して，提案法の MS（“HMM+MS”）は，自然音声の MS（“natural”）に接近している事が確認できる．

[*1] ポストフィルタ処理の一部は，HMM 音声合成ツールキット HTS (HMM-based Speech Synthesis System) ver. 2.3 beta [25] に搭載されている．

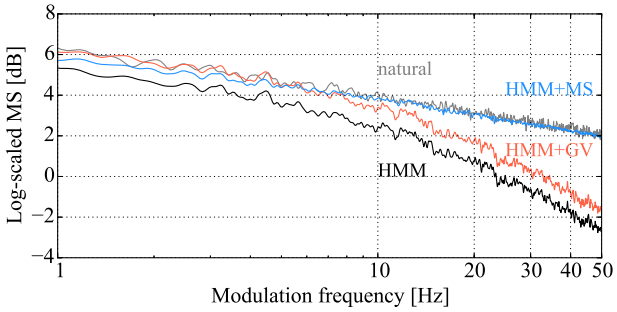


図 2 メルケプストラム系列の MS の例

Fig. 2 An example of the MS of the mel-cepstral coefficient sequences.

3.4 手法 3：MS 制約付きトラジェクトリ学習 [28], [29]

手法 2 の生成処理は，高音質の音声生成を可能にするものの生成処理を複雑にする[*2]ため，単遅延合成処理を必要とするシステムに不向きである．そこで，生成部に MS を組み込む代わりに，学習部に対して MS を導入する．各統計モデルは，静的・動的特徴量間の制約条件 $\boldsymbol{W}$ の下で系列をモデル化するトラジェクトリモデル [31] として表され，MS 尤度を考慮して学習される．

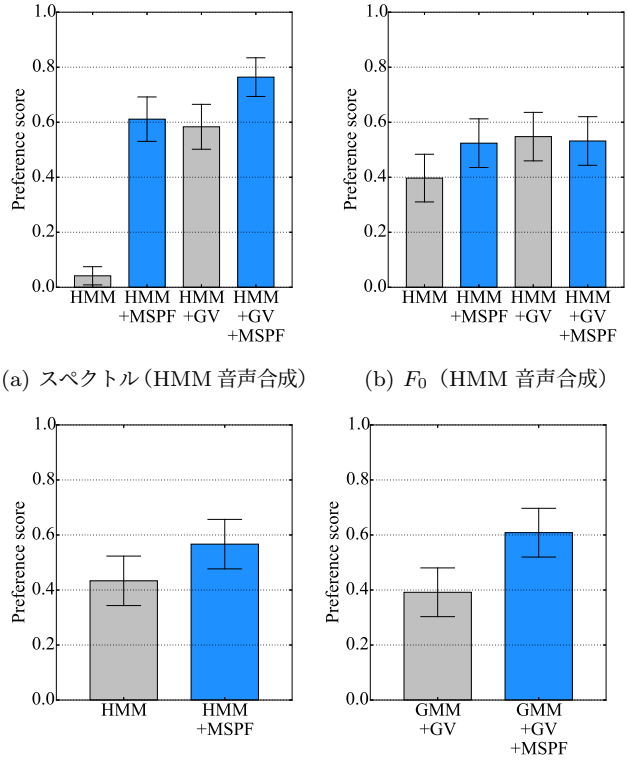
$$\boldsymbol{\lambda} = \mathrm{argmax}\, P\left(\boldsymbol{y}|\boldsymbol{W},\boldsymbol{X},\hat{\boldsymbol{q}},\boldsymbol{\lambda}\right) P\left(\boldsymbol{s}\left(\boldsymbol{y}\right)|\boldsymbol{X},\boldsymbol{\lambda},\boldsymbol{\lambda}_{\mathrm{s}}\right)^{\omega_{\mathrm{s}}} \quad (4)$$

MS 尤度 $P\left(\boldsymbol{s}\left(\boldsymbol{y}\right)|\boldsymbol{X},\boldsymbol{\lambda},\boldsymbol{\lambda}_{\mathrm{s}}\right)$ を考慮することで，統計モデルパラメータ $\boldsymbol{\lambda}$ は，式 (2) による生成パラメータ系列の MS $\boldsymbol{s}\left(\hat{\boldsymbol{y}}\right)$ が自然音声パラメータ系列の MS $\boldsymbol{s}\left(\boldsymbol{y}\right)$ に一致するように推定される．故に，生成部における MS 補償は不要であり，従来の高速生成処理が可能である．

4. 実験的評価

MS を補償するポストフィルタの主観的音質評価結果について示す．詳細な実験条件，客観評価結果，また，MS を考慮した音声パラメータ生成と MS 制約付きトラジェクトリ学習の評価結果については，[17], [23], [27], [29] を参照されたい．図 3 に，HMM 音声合成と GMM 声質変換におけるプリファレンス AB テストの評価結果を示す．“HMM” と “GMM” はそれぞれ，式 (2) による従来の生成法，“*+GV” は，生成時に GV を考慮したもの，“*+MSPF” は，ポストフィルタを適用した手法を示す．これらの結果から，ポストフィルタによる音質改善効果が得られ，特にスペクトルへの処理においてその効果が高いことが分かる．また，GV を考慮した場合でも改善効果があることが確認できる．図 4 に，メルケプストラム系列の例を示す．高次のメルケプストラム係数は，時間的に大きく振動する傾向にあり，過剰な平滑化の影響を強く受ける．そのため，フィルタ処理の影響は，高次メルケプストラムに対して特に大きいことが確認できる．

[*2] 従来の生成処理は，コレスキー分解 [21] や短遅延生成アルゴリズム [30] による高速生成が可能だが，手法 2 の生成処理は，勾配法による反復生成を必要とする．



(a) スペクトル（HMM 音声合成） (b) F_0（HMM 音声合成）

(c) 継続長（HMM 音声合成） (d) スペクトル（GMM 声質変換）

図 3 音質に関する主観評価結果（エラーバーは，95% 信頼区間）

Fig. 3 Preference scores on speech quality with the 95% confidence interval.

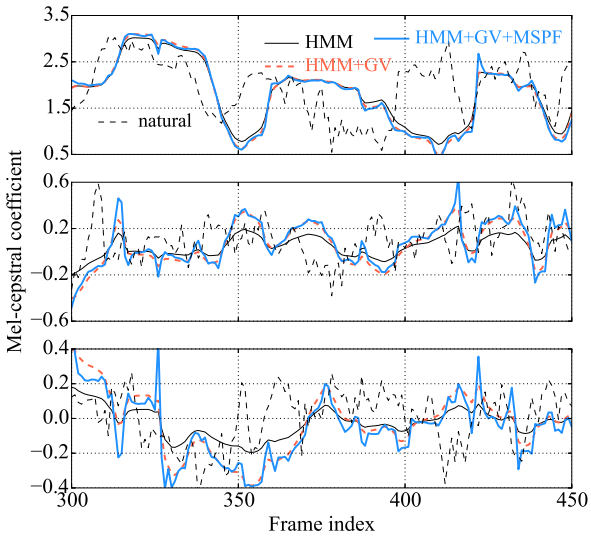


図 4 1 次，9 次，15 次のメルケプストラム系列の例

Fig. 4 Examples of the 1st, 9th, and 15th mel-cepstral coefficient sequences from above.

5. まとめ

本稿では，統計的パラメトリック音声合成の音質改善を目的として，変調スペクトル（MS）に基づく 3 手法（MS を補償するポストフィルタ，MS を考慮した音声パラメータ生成，MS 制約付きトラジェクトリ学習）を紹介した．

謝辞 本研究の一部は，JSPS 特別研究員奨励費 26·10354，JSPS 科研費 26280060，及び，頭脳循環を加速する若手研究者戦略的海外派遣プログラムの助成を受け実施した．

参考文献

[1] Y. Sagisaka. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. In *Proc. ICASSP*, pp. 679–682, New York, U.S.A., Apr. 1988.

[2] Y. Stylianou, O. Cappe, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 2, pp. 131–142, Mar. 1988.

[3] K. Tokuda, Y Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura. Speech synthesis based on hidden Markov models. *Proceedings of the IEEE*, Vol. 101, No. 5, pp. 1234–1252, 2013.

[4] T. Toda, A. W. Black, and K. Tokuda. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.

[5] H. Zen and A. Senior. Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In *Proc. ICASSP*, pp. 3872–3876, Florence, Italy, May 2014.

[6] K. Oura, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda. Tying covariance matrices to reduce the footprint of HMM-based speech synthesis systems. In *Proc. INTERSPEECH*, pp. 1759–1762, Brighton, U. K., 2009.

[7] J. Yamagishi and T. Kobayashi. Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE Trans., Inf. and Syst.*, Vol. E90-D, No. 2, pp. 533–543, 2007.

[8] J. Yamagishi, C. Veaux, S. King, and S. Renals. Speech synthesis technologies for individuals with vocal diabilities: Voice banking and reconstruction. *Acoust. Sci. technol.*, Vol. 33, pp. 1–5, 2012.

[9] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation. *IEICE Trans. on Inf. and Syst.*, Vol. E97-D, No. 6, pp. 1429–1437, Jun. 2014.

[10] S. Aryal and R. G.-Osuna. Can voice conversion be used to reduce non-native accents? In *Proc. ICASSP*, pp. 7929–7933, Florence, Italy, May 2014.

[11] 高道慎之介, 大島悠司, 戸田智基, Neubig Graham, Sakti Sakriani, 中村哲. 日本人英語のための音声合成技術を用いた英語学習支援の検討. 教育システム情報学会研究報告, Vol. 29, No. 5, pp. 111–116, Jan. 2015.

[12] K. Shirota, K. Nakamura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda. Integration of speaker and pitch adaptive training for HMM-based singing voice synthesis. In *Proc. ICASSP*, pp. 2578–2582, Florence, Italy, May 2014.

[13] K. Kobayashi, T. Toda, H. Doi, T. Nakano, M. Goto, G. Neubig, S. Sakti, and S. Nakamura. Voice timbre control based on perceived age in singing voice conversion. *IEICE Trans. on Inf. and Syst.*, Vol. E97-D, No. 6, pp. 1419–1428, Jun. 2014.

[14] S. King and V. Karaiskos. The blizzard challenge 2011. In *Proc. Blizzard Challenge workshop*, Turin, Italy, Sept. 2011.

[15] Y. Stylianou. Voice transformation: A survey. In *Proc. ICASSP*, pp. 3585–3588, Taipei, Taiwan, Apr. 2009.

[16] H. Zen, K. Tokuda, and A. Black. Statistical parametric speech synthesis. *Speech Commun.*, Vol. 51, No. 11, pp. 1039–1064, 2009.

[17] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. A postfilter to modify modulation spectrum in HMM-based speech synthesis. In *Proc. ICASSP*, pp. 290–294, Florence, Italy, May 2014.

[18] R. Drullman, J .M. Festen, and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. of America*, Vol. 95, pp. 2670–2680, 1994.

[19] S. Thomas, S. Ganapathy, and H. Hermansky. Phoneme recgnition usng spectral envelop and modulation frequency features. In *Proc. ICASSP*, pp. 4453–4456, Taipei, Taiwan, April 2009.

[20] Z. Wu, X. Xiao, E. S. Chng, and H. Li. Synthetic speech detection using temporal modulation feature. In *Proc. ICASSP*, pp. 7234–7238, Vancouver, Canada, May. 2013.

[21] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP*, pp. 1315–1318, Istanbul, Turkey, June 2000.

[22] T. Toda and K. Tokuda. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Trans.*, Vol. E90-D, No. 5, pp. 816–824, 2007.

[23] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura. Modulation spectrum-based post-filter for GMM-based voice conversion. In *Proc. APSIPA ASC*, Siem Reap, Cambodia, Dec. 2014.

[24] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura. Modified modulation spectrum-based post-filter for HMM-based speech synthesis. In *Proc. GlobalSIP*, pp. 710–714, Atlanta, United States, Dec. 2014.

[25] HMM-based speech synthesis system (HTS) http://hts.sp.nitech.ac.jp/.

[26] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura. Parameter generation algorithm considering modulation spectrum for HMM-based speech synthesis. In *Proc. ICASSP*, Brisbane, Australia, Apr. 2015.

[27] 高道慎之介, 戸田智基, A. W. Black, 中村哲. 統計的パラメトリック音声合成のための変調スペクトルを考慮した音声パラメータ生成アルゴリズム. 情報処理学会研究報告, Vol. 2015-SLP-105, No. 1, pp. 1–6, Feb. 2015.

[28] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura. Modulation spectrum-constrained trajectory training algorithm for GMM-based voice conversion. In *Proc. ICASSP*, Brisbane, Australia, Apr. 2015.

[29] 高道慎之介, 戸田智基, A. W. Black, 中村哲. 統計的パラメトリック音声合成のための変調スペクトル制約付きトラジェクトリ学習アルゴリズム. 電子情報通信学会技術研究報告, Vol. SP2015-03, pp. 31–36, Mar. 2015.

[30] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano. Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory. In *Proc. INTERSPEECH*, pp. 1076–1079, Brisbane, Australia, Sep. 2008.

[31] H. Zen, K. Tokuda, and T. Kitamura. Refomulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. *Computer Speech and Language*, Vol. 21, No. 1, pp. 153–173, Jan. 2007.