論文 / 著書情報
Article / Book Information

論題(和文)	系列内変動を考慮したガウス過程回帰に基づく音声パラメータ生成
Title(English)	
著者(和文)	郡山知樹, 能勢 隆, 小林隆夫
Authors(English)	Tomoki Koriyama, Takashi Nose, Takao Kobayashi
出典(和文)	日本音響学会2014年春季研究発表会講演論文集, Vol. , No. , pp. 355-356
Citation(English)	, Vol. , No. , pp. 355-356
発行日 / Pub. date	2014, 3

系列内変動を考慮したガウス過程回帰に基づく音声パラメータ生成*

◎郡山 知樹 (東工大)，能勢 隆 (東北大)，小林 隆夫 (東工大)

1 はじめに

高い自然性と柔軟性をもつ新たな統計的音声合成を目標に，我々はこれまでにガウス過程回帰に基づく音声合成手法を提案した [1, 2]．ガウス過程回帰を用いることで合成音声の音声パラメータの予測分布を推定することが可能であるが，その予測分布から，合成に用いる音声パラメータを生成することに関しては十分に検討を行っておらず，簡単な方法として予測分布の平均系列を用いる方法を用いていた．しかし，その予測平均により合成された音声は過剰平滑化によりこもった音に聞こえてしまうことがある．そこで本稿ではこの問題を避けるため，HMM 音声合成や声質変換において有効性が示されている系列内変動 (GV) による制約 [3] を用いた手法の検討を行う．

2 ガウス過程回帰に基づく音声合成

ガウス過程回帰 (GPR)[4] に基づく音声合成では，入力テキストとアクセント情報などから得られるフレームレベルのコンテキスト情報を入力変数 $\mathbf{x}_n$，フレームレベルの音響特徴量を出力変数 y_n とする回帰モデルを考える．GPR ではフレーム間の相関関係をカーネル関数 $k(\mathbf{x}_m, \mathbf{x}_n)$ で表すことによって，学習データに含まれない未知の入力変数に対応する音響特徴量の確率分布を予測することが可能であり，その予測分布を用いて音声を合成する．

Fig. 1 に GPR に基づく音声合成の流れを示す．音響特徴量のモデル化を次元毎に独立に行う．学習時には，まずハイパーパラメータの最適化を行う．ここでハイパーパラメータとは，フレーム間の相関関係を表すカーネル関数のパラメータおよびノイズを表すパラメータであり，これらの最適化を行うことで，データに適したカーネル関数を予測分布の計算に使用することができる．ハイパーパラメータの最適化には一般化 EM アルゴリズムが適用可能であることを文献 [2] で示している．学習時にはまた，学習データのフレーム間相関を示すグラム行列など，合成時の入力テキストに依存しないパラメータの計算を行う．

合成時には，以下の式を用いて合成音声の音響特徴量系列 $\mathbf{y}_T$ の予測分布を求める．

$$p(\mathbf{y}_T|\mathbf{y}_N) = \mathcal{N}(\mathbf{y}_T; \boldsymbol{\mu}_{\mathbf{y}_T|\mathbf{y}_N}, \boldsymbol{\Sigma}_{\mathbf{y}_T|\mathbf{y}_N}) \tag{1}$$

$$\boldsymbol{\mu}_{\mathbf{y}_T|\mathbf{y}_N} = \mathbf{K}_{TN}[\mathbf{K}_N + \sigma_\nu^2\mathbf{I}]^{-1}\mathbf{y}_N \tag{2}$$

$$\boldsymbol{\Sigma}_{\mathbf{y}_T|\mathbf{y}_N} = \mathbf{K}_T + \mathbf{K}_{TN}[\mathbf{K}_N + \sigma^2\mathbf{I}]^{-1}\mathbf{K}_{NT} + \sigma_\nu^2\mathbf{I} \tag{3}$$

ここで，$\mathbf{K}_N$，$\mathbf{K}_T$ はそれぞれ学習データ内，合成データ内のフレーム間相関を，$\mathbf{K}_{NT}$，$\mathbf{K}_{TN}$ は学習データ・合成データ間のフレーム間相関を表すグラム行列である．また，σ_ν^2 はノイズを表すパラメータであり，$\mathbf{y}_N$ は学習データ全体の音響特徴量系列である．逆行列を求めるには膨大な計算量が必要になるので，実現可能な計算量のために partially independent conditional(PIC) 近似 [5] を使用している [1].

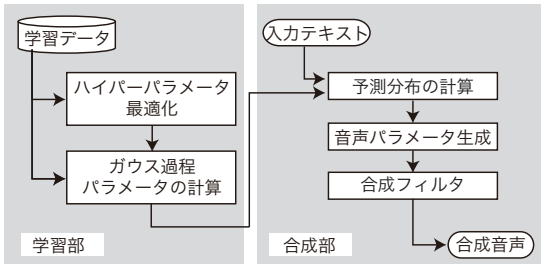

Fig. 1 ガウス過程回帰に基づく音声合成

最後に，GPR により求めた予測分布を用いて音声を合成する．具体的には，予測分布から音声パラメータ系列を生成し，その系列から合成フィルタを用いて音声を合成する．本稿ではこの音声パラメータ生成に GV を考慮した手法を導入する．

3 GV を考慮した音声パラメータ生成

GV は発話単位の特徴量であり，以下のように定義される．

$$v(\mathbf{y}_T) = \frac{1}{T}\sum_{t=1}^{T}\big(y_t - m(\mathbf{y}_T)\big)^2 \tag{4}$$

$$m(\mathbf{y}_T) = \frac{1}{T}\sum_{t=1}^{T} y_t \tag{5}$$

ここで T は発話のフレーム数であり．GV を制約条件に用いた場合の尤度は以下のように定義される．

$$\mathcal{L}_{\mathrm{GV}} = p(\mathbf{y}_T|\mathbf{y}_N)^\omega p\big(v(\mathbf{y}_T)|\mu_{\mathrm{GV}}, \sigma_{\mathrm{GV}}^2\big) \tag{6}$$

μ_{GV}，σ_{GV}^2 はそれぞれ学習データにおける発話毎の GV の平均，分散を表す．パラメータ ω はガウス過程による予測尤度と GV 尤度との重みを調節するパラメータである．

尤度 $\mathcal{L}_{\mathrm{GV}}$ を最大化することによって音声パラメータ系列を生成する．実際には，解析的に最適なパラメータ系列を求めることはできないので，以下に示す勾配を用いて最急降下法により最適なパラメータ系列を求める．

$$\frac{\partial\mathcal{L}_{\mathrm{GV}}}{\partial\mathbf{y}_T} = -\omega\boldsymbol{\Sigma}_{\mathbf{y}_T|\mathbf{y}_N}^{-1}(\mathbf{y}_T - \boldsymbol{\mu}_{\mathbf{y}_T|\mathbf{y}_N}) - \frac{2}{T}\cdot\frac{v(\mathbf{y}_T) - \mu_{\mathrm{GV}}}{\sigma_{\mathrm{GV}}^2}\big(\mathbf{y}_T - m(\mathbf{y}_T)\cdot\mathbf{1}\big) \tag{7}$$

4 実験

4.1 実験条件

音声データベースとして女性話者一人により発話された ATR 音素バランス文 503 文を使用し，学習

*Speech parameter generation based on Gaussian process regression using global variance. by KORIYAMA, Tomoki (Tokyo Institute of Technology), NOSE, Takashi (Tohoku University), and KOBAYASHI, Takao (Tokyo Institute of Technology)

Table 1　合成音声の原音声に対するメルケプストラム距離 [dB]

	HMM	GPR
GV なし	5.38	5.03
GV あり	5.95	5.55

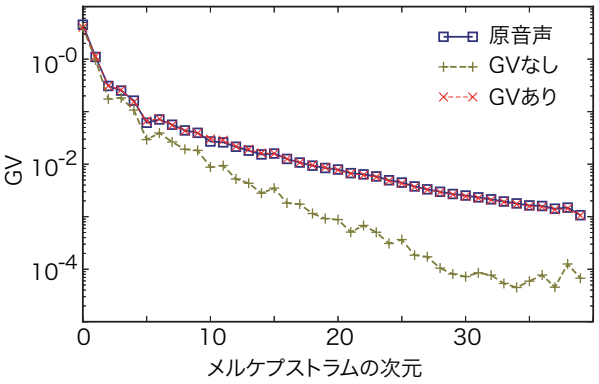

Fig. 2　合成音声における GV の全発話の平均

データには 450 文を，テストデータには学習データに含まれない 53 文を用いた．周波数 16kHz でサンプリングされた音声に対し，5 ミリ秒毎に STRAIGHT を用いてスペクトル包絡を抽出し，それをもとに得られた 0 次から 39 次のメルケプストラムを音響特徴量とした．音響特徴量は平均 0，分散 1 となるように正規化を行いモデル化は次元毎に行った．GV 重みを表すパラメータ ω は $v(\mathbf{y}_T)$ と $\mathbf{y}_T$ の次元比 $1/T$ とした．GPR のカーネル関数およびハイパーパラメータの初期値，PIC 近似の設定には [2] と同様のものを用いた．

比較手法には，GV を考慮しない場合として予測分布の平均を生成系列とする方法を使用した．また，従来手法として隠れマルコフモデル (HMM) に基づく手法 (HMM 音声合成) を用いた．HMM 音声合成では 5 状態の left-to-right スキップなし隠れセミマルコフモデル (HSMM) によって音声のモデル化を行った．HSMM の各状態の出力分布は対角共分散行列を持つ単一ガウス分布とし，Δ，Δ^2 の動的特徴量を含む特徴ベクトルを音響特徴量として用いた．コンテキスト決定木学習時には，トライフォンをコンテキストとして使用した．誤差最小化学習 (MGE) による最適化を行う手法 [6] を用いた．

主観評価実験における被験者は 6 人で，各被験者は合成音声の自然性を 5 段階 (1:bad〜5:excellent) で評価した．各被験者に対し 15 文章をランダムに選択し，その文章をそれぞれの手法で合成した音声を用いた．本実験では，メルケプストラムのみモデル化を行い，音素継続長，F0，非周期性指標には原音声のものを用いた．

4.2　結果

合成音声のスペクトル歪を Table 1 に示す．表中の GV なし・GV ありとはそれぞれ GV を考慮しない・した場合を表す．HMM 音声合成において GV を考慮した場合と比較すると，GPR において GV を考慮した場合はスペクトル歪が小さくなっていることが分かる．また，合成音声の音声パラメータの全発話に対する GV の平均を Fig. 2 に示す．結果から，GV を考慮することで原音声に近い GV を再現できていることが確認できる．

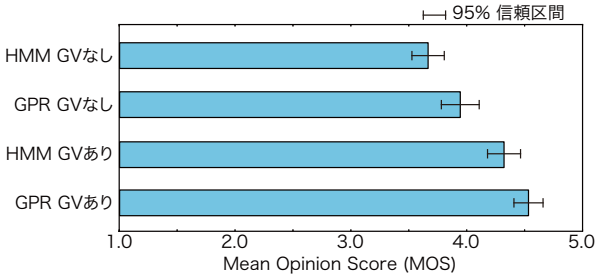

Fig. 3　合成音声の自然性に対する主観評価

Fig. 3 に主観評価実験の結果を示す．スコアが通常のテキスト音声合成に比べ比較的高めなのは，本実験では韻律の特徴量に原音声のものを用いたためである．結果から GPR に基づく音声合成においても GV を用いることで自然性が向上していることを確認できる．さらに，GPR において GV を考慮することで HMM 音声合成に比べスコアが高くなり，その差は有意水準 $\alpha = 0.05$ で有意であった．

5　おわりに

本稿ではガウス過程回帰に基づく音声合成の枠組みにおいて，系列内変動を考慮した音声パラメータの生成手法を検討した．主観評価実験の結果から，GV を考慮することで自然性が向上することを示し，また，従来の HMM に基づく手法よりも高い自然性が得られることがわかった．今後は，F0 および音素継続長などの韻律情報に対するモデル化手法を検討する予定である．

謝辞　本研究の一部は，日本学術振興会科学研究費補助金 (課題番号 24300071, 25540065, 25・8776) の助成を得た．

参考文献

[1] 郡山他，“スパース近似と畳み込みカーネルを用いたガウス過程回帰に基づく音声合成，” 音講論 (秋)，2-7-12，pp.311–312，2013.

[2] 郡山他，“ガウス過程回帰に基づく音声合成におけるハイパーパラメータ最適化の検討，” 信学技報，vol.113，no.404，SP2013-99，pp.19–24，2014.

[3] Toda et al., “A Speech parameter generation algorithm considering global variance for HMM-based speech synthesis,” IEICE Trans. Inf. & Syst., E90-D, no.5, pp.816–824, 2006.

[4] C.E. Rasmussen and C.K.I. Williams, *Gaussian processes for machine learning*, MIT press, 2006.

[5] E. Snelson and Z. Ghahramani, “Local and global sparse Gaussian process approximations,” Proc. AISTATS, pp.524–531, 2007.

[6] Y.J. Wu and R.H. Wang, “Minimum generation error training for HMM-based speech synthesis,” *Proc. ICASSP*, pp. 889–892, 2006.