論文 / 著書情報
Article / Book Information

論題(和文)	ガウス過程回帰に基づく音声合成システムの検討
Title(English)	
著者(和文)	郡山知樹, 小林隆夫
Authors(English)	Tomoki Koriyama, Takao Kobayashi
出典(和文)	日本音響学会2015年春季研究発表会講演論文集, Vol. , No. , pp. 269-270
Citation(English)	, Vol. , No. , pp. 269-270
発行日 / Pub. date	2015, 3

ガウス過程回帰に基づく音声合成システムの検討*

◎郡山知樹, 小林隆夫 (東工大)

1 はじめに

我々はこれまでに，ガウス過程回帰 (GPR) に基づくフレームレベル音響モデリングを用いた，スペクトル特徴量 [1] および F0 パタン [2] の予測モデルを提案した．本稿ではさらに GPR に基づく音素継続長の予測モデルを加えたガウス過程回帰音声合成 (GPR 音声合成) システムの実験的評価を行う．また，本研究では合成音声の自然性の向上に向け，滑らかに変化する音声パラメータ生成のための動的特徴量の導入を検討する．

2 ガウス過程回帰音声合成

本研究で提案するガウス過程回帰音声合成 (GPR 音声合成) では，Table 1 に示すメルケプストラムモデル，有声／無声モデル，F0 モデル，非周期性指標モデル，音素継続長モデルの 5 つのモデルを使用し，それぞれ独立にモデル化，音声パラメータ生成を行う．文献 [2] に示すように，有声／無声モデルにはガウス過程分類の枠組みを使用し，F0 は有声フレームに対してのみモデル化を行う．また，音素継続長は音素単位の特徴量であるため，音素単位でモデル化する．

ここで，メルケプストラムと非周期性指標は多次元ベクトルである．文献 [1] では，メルケプストラムの各次元を別々にモデル化していたが，次元毎の音声パラメータ生成には計算時間が次元数に比例してしまうという問題がある．そこで，本研究では以下に示す多次元ベクトルの同時モデル化手法を用いる．

まず次元 d $(d = 1, \ldots, D)$ に対し，学習データの平均 μ_d と分散 σ_d^2 で正規化した，サンプル n の音響特徴量 $\bar{y}_n^{(d)} = (y_n^{(d)} - \mu_d)/\sigma_d$ の同時分布を考える．同時分布がすべての次元に共通のグラム行列 $\mathbf{K}_N$ を用いて

$$p\left(\bar{\mathbf{y}}_N^{(d)}\right) = \mathcal{N}\left(\bar{\mathbf{y}}_N^{(d)}; \mathbf{0}, \mathbf{K}_N + \sigma_\nu^2 \mathbf{I}\right) \tag{1}$$

$$\bar{\mathbf{y}}_N^{(d)} = \left[\bar{y}_1^{(d)}, \quad \ldots, \quad \bar{y}_N^{(d)}\right]^\top \tag{2}$$

で与えられ，各次元の分布が独立であると仮定する．このとき，合成文の音響特徴量系列 $\mathbf{Y}_T$ の予測分布は以下の式で求められる．

$$p(\mathbf{Y}_T|\mathbf{Y}_N) = \mathcal{MN}(\mathbf{Y}_T; \mathbf{M}, \mathbf{\Sigma}, \mathbf{V}) \tag{3}$$

$$\mathbf{M} = \mathbf{K}_{TN}\left(\mathbf{K}_N + \sigma_\nu^2 \mathbf{I}\right)^{-1}\left(\mathbf{Y}_N - \mathbf{1}\mathbf{b}^\top\right) + \mathbf{1}\mathbf{b}^\top \tag{4}$$

$$\mathbf{\Sigma} = \mathbf{K}_T - \mathbf{K}_{TN}\left(\mathbf{K}_N + \sigma_\nu^2 \mathbf{I}\right)^{-1}\mathbf{K}_{NT} + \sigma_\nu^2 \mathbf{I} \tag{5}$$

ただし，$\mathcal{MN}(\cdot)$ は行列変量正規分布 [3] であり，$\mathbf{Y}_N = \left[\mathbf{y}_N^{(1)}, \quad \ldots, \quad \mathbf{y}_N^{(D)}\right]$，$\mathbf{V} = \mathrm{diag}\left[\sigma_1^2, \ldots, \sigma_D^2\right]$，$\mathbf{b} = \left[\mu_1, \quad \ldots, \quad \mu_D\right]^\top$ である．また，$\mathbf{K}_N$，$\mathbf{K}_T$ 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．なお計算量削減のため各グラム行列は部分独立条件 (PIC) 近似 [4] によるブロック近似を行う．

Table 1 ガウス過程回帰音声合成におけるモデル

モデル	出力変数	予測	単位
メルケプストラム	多次元連続	回帰	フレーム
有声／無声	二値	クラス分類	フレーム
F0	一次元連続	回帰	フレーム
非周期性指標	多次元連続	回帰	フレーム
音素継続長	一次元連続	回帰	音素

3 動的特徴量を用いた音声パラメータ生成

GPR 音声合成ではフレームレベルの音響特徴量を直接予測できるが，計算量削減のための PIC 近似によるブロック近似のため，ブロックの境界において音声パラメータに不連続性が生じてしまうという問題がある．そこで，本研究では HMM 音声合成において利用されている動的特徴量を用いた音声パラメータ生成手法を GPR 音声合成の枠組みに導入する．

静的特徴量 $\mathbf{C}$ に対し，動的特徴量を含む観測系列が

$$\begin{aligned}\mathbf{Y} &= \left[\mathbf{C} \quad \Delta\mathbf{C} \quad \Delta^2\mathbf{C}\right] \\ &= \left[\mathbf{W}_0\mathbf{C} \quad \mathbf{W}_1\mathbf{C} \quad \mathbf{W}_2\mathbf{C}\right]\end{aligned} \tag{6}$$

で表されるとする．観測系列 $\mathbf{Y}$ に対する分布が行列変量正規分布

$$p(\mathbf{Y}) = \mathcal{MN}(\mathbf{Y}; \mathbf{M}, \mathbf{\Sigma}, \mathbf{V}) \tag{7}$$

$$\mathbf{M} = \left[\mathbf{M}_0 \quad \mathbf{M}_1 \quad \mathbf{M}_2\right] \tag{8}$$

$$\mathbf{V} = \mathrm{diag}\left[\mathbf{V}_0, \quad \mathbf{V}_1, \quad \mathbf{V}_2\right] \tag{9}$$

で与えられるとき，静的特徴量の確率分布は

$$p(\mathbf{C}) = \frac{1}{K}\exp\left(-\frac{1}{2}\sum_{i=0}^{2}\mathrm{Tr}\left[\mathbf{V}_i^{-1}(\mathbf{W}_i\mathbf{C} - \mathbf{M}_i)^\top \mathbf{\Sigma}^{-1}(\mathbf{W}_i\mathbf{C} - \mathbf{M}_i)\right]\right) \tag{10}$$

で表される．ただし，K は正規化定数である．

尤度最大化基準に基づく音声パラメータ生成時には，勾配

$$\frac{\partial \log p(\mathbf{C})}{\partial \mathbf{C}} = -\sum_{i=0}^{2}\mathbf{W}_i^\top \mathbf{\Sigma}^{-1}(\mathbf{W}_i\mathbf{C} - \mathbf{M}_i)\mathbf{V}_i^{-1} \tag{11}$$

を用いて，勾配法により最大値を求めるか，式 (11) が 0 になる解を直接求める．式 (11) を 0 にする解を直接求めるには $\mathcal{O}(D^3T^3)$ の計算量が必要であるため，本研究では一次元の対数 F0 の生成には直接解を，多次元のメルケプストラムおよび非周期性指標の生成には勾配法を用いる．

4 実験

4.1 実験条件

実験には ATR 日本語音声データベースセット B に含まれる女声話者，男声話者，各 2 名の計 4 名 (FKS,

*A study on speech synthesis system based on Gaussian process regression, by KORIYAMA, Tomoki and KOBAYASHI, Takao (Tokyo Institute of Technology)

Table 2　合成音声の原音声に対する音響特徴量歪

(a) メルケプストラム距離 [dB]

話者＼手法	HMM	GPR-STA	GPR-DYN
FKS	4.80	4.68	4.55
FTK	4.87	4.69	4.58
MMY	5.36	5.17	5.04
MHT	4.41	4.48	4.35
AVE	4.86	4.75	4.63

(b) 対数 F0 の RMS 誤差 [cent]

話者＼手法	HMM	GPR-STA	GPR-DYN
FKS	193	170	163
FTK	224	208	202
MMY	220	181	175
MHT	184	156	152
AVE	205	179	173

(c) 音素継続長の RMS 誤差 [ms]

話者＼手法	HMM	GPR-STA/GPR-DYN
FKS	22.5	20.8
FTK	23.6	21.2
MMY	20.5	18.8
MHT	23.5	21.7
AVE	22.5	20.6

FTK, MMY, MHT) の音声を用いた．学習データには 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，STRAIGHT を用いて基本周波数およびスペクトル包絡，非周期性指標を抽出した．スペクトル特徴量には STRAIGHT スペクトルから得られる 0〜39 次のメルケプストラム，対数 F0，5 次元の非周期性指標を音響特徴量として使用した．

ガウス過程回帰・分類に用いる PIC 近似 [4] におけるブロックの最大フレーム数は 1024 とし，学習データに含まれるフレーム全体からランダムに選択した 1024 フレームを疑似データセットとした．このとき，PIC 近似におけるブロックの決定には HMM 音声合成の枠組みで得られるコンテキスト決定木を使用した．

比較手法には HMM 音声合成を使用し，モデルは 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) とした．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，メルケプストラム，対数 F0，非周期性指標とそれらの Δ，Δ^2 の動的特徴量を含む 138 次元の特徴ベクトルを音響特徴量として用いた．

4.2　結果

従来手法の HMM 音声合成 (HMM)，静的特徴量のみを使用した GPR 音声合成 (GPR-STA)，メルケプストラム，対数 F0，非周期性指標に動的特徴量を導入した GPR 音声合成 (GPR-DYN) それぞれの原音声に対する合成音声の音響特徴量歪を Table 2 に示す．表中の値は 4 話者の平均を示している．表から，提案法の GPR-STA は HMM 音声合成に比べメルケプストラム距離，対数 F0 の RME 誤差，音素継続長の RMS 誤差において歪が小さくなる傾向にあり，動的特徴量を加えた GPR-DYN では歪がさらに小さくなっていることがわかる．

次に，合成音声の自然性を主観評価実験により比較した．主観評価実験における被験者は 7 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 10 文章を話者ごとにランダムに選択した．本実験では客観評価に使用した 3 手法 (HMM, GPR-STA, GPR-DYN) に加え，自然性の向上に大きく寄与している特徴量を調べるため，GPR-DYN の特徴量のうち，メルケプストラム，F0，音素継続長をそれぞれ HMM 音声合成で得られる音声パラメータにそれぞれ差し替えた H-MCEP, H-F0, H-DUR を比較手法とした．

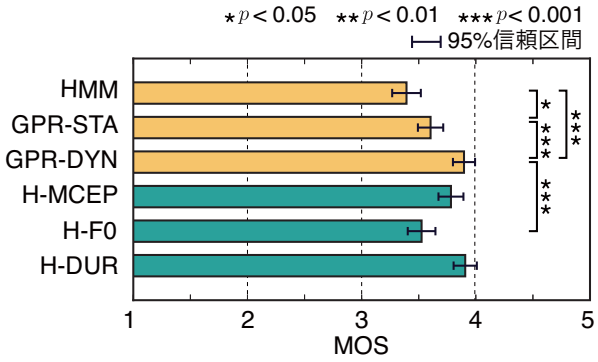

Fig. 1　合成音声の自然性に関する主観評価結果

Fig. 1 に結果を示す．HMM，GPR-STA，GPR-DYN を比較すると，GPR-STA は HMM よりスコアが高く，さらに動的特徴量を導入した GPR-DYN は GPR-STA より高いスコアを得た．3 手法の各手法間のスコアの差は有意水準 0.05 で有意であった．また，GPR-DYN とその音声パラメータの一部を変更した H-MCEP, H-F0, H-DUR とを比較すると，GPR-DYN は H-F0 に比べ有意にスコアが高く，このことから F0 モデルの自然性に向上に対する寄与が大きいことがわかる．

5　おわりに

本稿では，ガウス過程回帰 (GPR) を用いた音声合成システムを提案し，主観評価実験による評価を行った．結果として，GPR 音声合成を用いることで HMM 音声合成に比べ，自然性の高い音声が合成可能であり，さらに動的特徴量に基づく音声パラメータ生成を導入することで，自然性が向上することを示した．今後は合成に要する時間と性能の関係の評価や，HMM 音声合成以外の手法との比較，多様なスタイルの音声に対しても本手法を適用し評価を行う予定である．

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] T. Koriyama et al.,, “Statistical Parametric Speech Synthesis Based on Gaussian Process Regression,” IEEE J-STSP, (8)2, pp.173–183.

[2] 郡山 他，“ガウス過程回帰に基づく F0 パタン生成の検討，” 音講論 (秋)，2-7-8，pp.247–248，2014

[3] P. Dutilleul, “The MLE algorithm for the matrix normal distribution,” Journal of Statistical Computation and Simulation, vol.64, no.2, pp.105-123, 1999.

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