T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	ガウス過程回帰に基づく歌声合成の検討
Title(English)	
著者(和文)	郡山知樹, 岡野 祐紀, 小林隆夫
Authors(English)	Tomoki Koriyama, Yuuki Okano, Takao Kobayashi
出典(和文)	日本音響学会2017年秋季研究発表会講演論文集, Vol. , No. , pp. 295-296
Citation(English)	, Vol. , No. , pp. 295-296
発行日 / Pub. date	2017, 9

Powered by T2R2 (Science Tokyo Research Repository)

ガウス過程回帰に基づく歌声合成の検討*

○郡山知樹, △岡野祐紀, 小林隆夫 (東工大)

1 はじめに

統計モデルに基づく歌声合成システムは，実際の歌唱データを基に任意の楽曲の歌声を生成するシステムである．このシステムには歌唱者の歌唱パターンを学習し，より自然な歌声を生成することが期待される．これまでの研究では，統計的音声合成のシステムを歌声に適用した隠れマルコフモデル (HMM) に基づく歌声合成 [1] やディープニューラルネットワーク (DNN) に基づく歌声合成 [2, 3] が提案されている．一方で我々はガウス過程回帰 (GPR) に基づく音声合成を提案し，DNN 音声合成と同程度の自然性が実現可能であることを示している [4]．そこで本研究では，GPR に基づく音声合成の歌声合成へ適用について検討を行う．

2 GPR に基づく音響モデリング

GPR 音声合成 [4] の枠組みでは，フレームレベルの音響特徴量 y が，ガウス過程 $\mathcal{GP}(\cdot)$ に従う潜在関数 $f(\cdot)$ によって以下の式で生成されることを仮定する．

$$y = f(\mathbf{x}) + \epsilon \tag{1}$$

$$f \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{2}$$

ただし，ϵ は分散 σ_ν^2 に従うガウスノイズであり，$m(\mathbf{x})$, $k(\mathbf{x}, \mathbf{x}')$ はそれぞれフレームレベルのコンテキスト $\mathbf{x}$ に対する平均関数および共分散関数（カーネル関数）である．このとき，学習データ $(\mathbf{X}_N, \mathbf{y}_N)$ を $\mathbf{X}_N = \{\mathbf{x}_n\}_{n=1}^M$, $\mathbf{y}_N = [y_1, \ldots, y_N]^\top$ とすると，未知のデータ $(\mathbf{X}_T, \mathbf{y}_T)$ の予測分布は以下のように表される．

$$p(\mathbf{y}_T|\mathbf{y}_N) = \mathcal{N}(\mathbf{y}_T; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3}$$

$$\boldsymbol{\mu} = \mathbf{K}_{TN}\left(\mathbf{K}_N + \sigma_\nu^2\mathbf{I}\right)^{-1}(\mathbf{y}_N - \mathbf{m}_N) + \mathbf{m}_T \tag{4}$$

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

ただし，$\mathbf{m}_N = \left[m(\mathbf{x}_1), \ \ldots, \ m(\mathbf{x}_N)\right]^\top$ であり，$\mathbf{K}_N$，$\mathbf{K}_T$ 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それぞれ表すグラム行列である．実際に $\left(\mathbf{K}_N + \sigma_\nu^2\mathbf{I}\right)^{-1}$ を計算することは計算量の観点から困難であるため，各グラム行列に対して部分独立条件 (PIC) 近似 [5] によるブロック近似を行う．

3 GPR に基づく歌声合成

本研究では，2 節で述べた音響モデリング手法を歌声合成に適用する．具体的には，歌声合成に適したコンテキストの導入と音高との差分を考慮したモデリング手法を提案する．

Table 1 歌声合成のためのフレームレベルコンテキスト

相対位置
・{ 音素，音符，小節，フレーズ } の { 開始，終了 } からの位置
イベント特徴量
・音素名 one-hot ベクトル
・音素弁別特性 (13 次元バイナリベクトル)
・音符の音高
・音符の音高の差分
・音符の継続時間
・小節の継続時間

3.1 歌声合成のためのコンテキスト

文献 [6] に示すように，GPR 音声合成では音素や句の開始／終了位置で音響的なイベントが発生すると仮定し，イベントの発生時刻からのフレームの相対位置および，イベントの特徴量（母音／子音，アクセントの高／低など）をフレームレベルのコンテキストとして用いる．歌声合成では，楽譜情報を入力変数として考えるため，高低アクセントやアクセント句はコンテキストとして適切ではない．そこで本研究では表 1 に示すコンテキストを用いる．表中の音符，小節は歌声合成固有のコンテキストとなる．また本研究では，楽曲を休符のみを含む小節で区切った区間をフレーズとする．

3.2 音高との差分を考慮したモデリング

歌唱音声の基本周波数 f_o は，基本的に楽譜で表される音高にそって発声される．そこで HMM 歌声合成の枠組みでは，f_o の音高からの差分を学習パラメータに用いる音高正規化学習 [7] によって，f_o を効率的に学習するだけでなく，データベースに含まれない音高に対しても頑健な f_o 曲線の生成を可能にしている．一方で文献 [2] の DNN 歌声合成では，モデルパラメータに直接音高楽譜から得られた音符の音高と対数 f_o との差分を特徴量としてモデル化する枠組みが提案されている．

本研究においても，DNN 音声合成と同様に対数 f_o と音符の音高との差分をピッチ特徴量の静的特徴量として学習する．このとき休符部分の音高には隣接する音符の音高を線形補完したものを用いる．この手法は式 (2) におけるガウス過程の平均関数 $m(\mathbf{x})$ を，「音符の音高に対応する周波数」を出力する関数と見なすことができる．

---

*A Study on Singing Voice Synthesis Based on Gaussian Process Regression. by KORIYAMA, Tomoki, OKANO, Yuki, KOBAYASHI, Takao (Tokyo Institute of Technology)

Table 2　客観評価結果．MCEP: メルケプストラム距離 [dB]，f_o: 対数 f_o の RMS 誤差 [cent].

手法	MCEP	f_o
HMM	5.47	59.9
GPR	5.34	46.5

4　実験

4.1　実験条件

歌声データベースには，女性による童謡 30 曲，30 分 52 秒の無伴奏の歌声を使用した．それらのうち 25 曲，27 分 10 秒を学習データとして使用し，残りの 5 曲（10 フレーズ），3 分 42 秒を評価データとして使用した．サンプリングレート 16kHz の歌唱音声信号から，5ms ごとに STRAIGHT を用いて f_o，スペクトル包絡，非周期性指標を抽出し，音響特徴量として 0–39 次のメルケプストラム，対数 f_o，5 次元の非周期性指標，およびそれらの Δ, Δ^2 を使用した．コンテキストには MusicXML の楽譜データから自動で取得したものを使用した．フレームレベルコンテキストの次元は 261 であった．

GPR の部分独立条件 (PIC) 近似の補助点数は 1024，各ブロックの最大フレーム数は 1024 とし，ブロックの割り当てには HMM に基づく決定木を使用した．また，相対位置コンテキストとイベント特徴量コンテキストの類似性を表すカーネル関数には RBF カーネルを用いた．比較手法として HMM に基づく歌声合成を使用した．なお初期の検討のため，本研究では音素継続長のモデル化は行わず，原音声の音素境界を用いて歌声合成を行った．

4.2　結果

原音声と合成歌唱音声の音響特徴量歪を表 2 に示す．表から，GPR に基づく提案法では HMM に比べメルケプストラム距離，f_o の RMS 誤差がどちらも小さくなっていることがわかる．

また，生成された f_o の比較を行うため，童謡【かたつむり】の「つのだせやりだせ あたまだせ」の合成歌唱音声の f_o を図 1 に示す．図中の垂直方向の目盛り線は楽譜情報によって与えられる音符の境界を，水平方向の目盛り線は平均律における音高を，それぞれ表している．図から，HMM では音符内で f_o がほぼ一定であるのに対し，GPR では音符内で f_o が弧を描いて不安定な系列になっていることがわかる．

このような f_o 系列となっている原因の一つとして，外れ値の影響が考えられる．本研究の枠組みでは，音符毎に対数 f_o と音高との差分を計算してピッチ特徴量として用いている．このとき，音符の中心付近ではピッチ特徴量は 0 に近く境界付近で音響特徴量は絶対値の大きな値となる．そのため，この境界付近の特徴量が外れ値となって不安定な系列を生成している可能性がある．

この音符境界の特徴量の影響を抑えるための解決法としては，明示的な平均関数 $m(\cdot)$ を導入する方法が考えられる．平均関数を用いて音符系列から予想

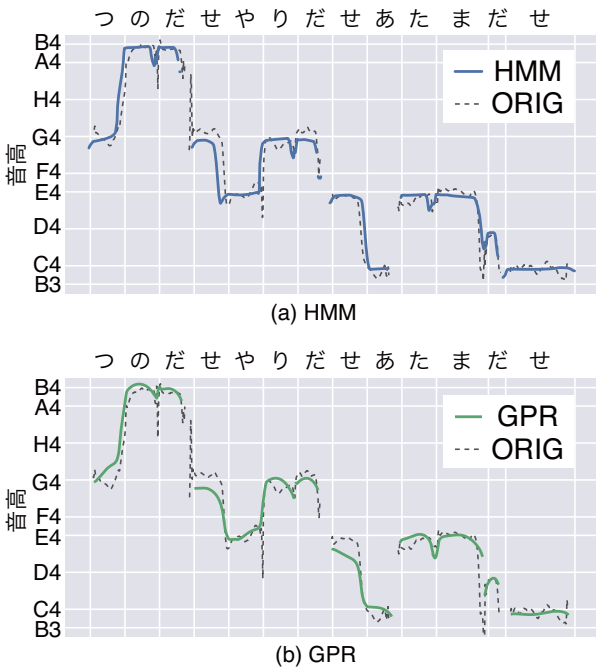

Fig. 1　合成歌唱音声における f_o 系列の比較

されるピッチカーブを求められれば，それからの差分はスムーズな系列になり，外れ値の影響を受けにくくなることが期待できる．

5　おわりに

本研究では，GPR に基づく統計的歌声合成の枠組みとして，GPR 音声合成に歌声特有のコンテキストの導入および，音高との差分を考慮したモデル化の導入を行った．結果として，HMM に基づく音声合成に比べスペクトル歪および f_o 歪が小さくなることを示した．今後は主観評価実験による HMM や DNN に基づく手法との比較や，考察で述べたような f_o 系列が不安定になってしまう問題の解決が必要である．

謝辞　本研究の一部は JSPS 科研費 JP15H02724 の助成を受けた．

参考文献

[1] K. Oura et al., “Recent Development of the HMM-based Singing Voice Synthesis System — Sinsy,” Proc. of Speech Synthesis Workshop, pp. 211–216, 2010.
[2] 西村 他，“Deep Neural Network に基づく歌声合成の検討,” 音講論 (春)，1-1-2, pp. 213–214，2016.
[3] 本郷 他，“ディープニューラルネットワークに基づく歌声合成の検討,” 音講論 (春)，1-R-24，pp. 295–296，2016.
[4] 郡山 他，“ガウス過程回帰に基づく音声合成システムの評価,” 音講論 (秋)，3-1-3，pp. 235–236，2015.
[5] E. Snelson & Z. Ghahramani, “Local and global sparse Gaussian process approximations,” Proc. AISTATS, pp.524–531, 2007.
[6] 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.
[7] 間瀬 他，“音高正規化学習を用いた HMM 歌声合成の検討,” 音講論 (秋)，1-8-20，pp. 283–284，2011.