論文 / 著書情報
Article / Book Information

論題(和文)	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
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2013年秋期研究発表会講演論文集, Vol. , No. , pp. 311-312
Citation(English)	, Vol. , No. , pp. 311-312
発行日 / Pub. date	2013, 9

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*

☆郡山 知樹，能勢 隆，小林 隆夫 (東工大)

1 はじめに

新たな統計的音声合成の枠組みとして，我々はガウス過程回帰に基づくフレームレベルの音声の合成手法を提案している [1]．ガウス過程回帰 [2] はノンパラメトリックベイズモデルとして知られており，学習データ量に応じて柔軟にモデル化を行うことが期待できる．文献 [1] では孤立音素単位に対する合成を行ったが，本手法を発話単位での連続音声の合成に直接適用することは困難である．例えば，学習データ量が増加すると計算量が非常に大きくなってしまう．また，音素の境界を適切にモデル化する必要がある．そこで本稿では，連続音声の合成を行うために，スパース行列を用いた近似と畳み込みカーネルを導入する．

2 ガウス過程回帰に基づく音声合成

ガウス過程に基づく回帰 [2] では，学習データを $\mathcal{D} = \{(\mathbf{x}_n, y_n)|n = 1, \ldots, N\}$，評価データを $\mathcal{D}_T = \{(\mathbf{x}_t, y_t)|t = 1, \ldots, T\}$ としたとき，平均 0 に正規化された学習データ $\mathbf{y} = [y_1, \ldots, y_N]^\top$ および評価データ $\mathbf{y}_T = [y_1, \ldots, y_T]^\top$ の同時分布は以下の式で表される．

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_T \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{K}_{N+T} + \sigma^2\mathbf{I}\right) \tag{1}$$

$$\mathbf{K}_{N+T} = \begin{bmatrix} \mathbf{K}_N & \mathbf{K}_{NT} \\ \mathbf{K}_{TN} & \mathbf{K}_T \end{bmatrix} \tag{2}$$

ただし，σ^2 はノイズの分散である．また，$\mathbf{K}_N$，$\mathbf{K}_T$ は学習データおよび評価データそれぞれのフレーム間の共分散行列であり，$\mathbf{K}_{NT}$，$\mathbf{K}_{TN}$ は学習・評価データ間の共分散行列である．共分散行列の要素はカーネル関数 $k(\mathbf{x}_m, \mathbf{x}_n)$ で与えられる．このとき，評価データの予測分布はガウス分布に従い，その平均系列は

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

で求められる．ガウス過程回帰に基づく音声合成では，フレームレベルのコンテキストを入力変数とし対応するフレームの音響特徴量を出力変数とする回帰モデルを使用する．

3 スパース近似を用いたガウス過程

連続音声の合成における問題点の一つは，学習データが増加すると計算量が非常に大きくなることである．具体的には，式 (3) の逆行列の計算に $\mathcal{O}(N^3)$ の計算量が必要になる．そこで，ガウス過程の近似手法として局所 GP[3, 4] および PIC(partially independent conditional) 近似 [4] を導入する．

局所 GP はデータ全体を複数のブロックに分割し，共分散行列を以下のようにブロック対角で近似する手法である．

$$\begin{aligned} \mathbf{K}_{N+T}^{\mathrm{LOCAL}} &= \mathrm{blkdiag}\left[\mathbf{K}_{N+T}\right] \\ &= \mathrm{diag}\left[\mathbf{K}_{B_1}, \mathbf{K}_{B_2}, \ldots, \mathbf{K}_{B_S}\right] \end{aligned} \tag{4}$$

ここで，$\mathbf{K}_{B_s}$ はブロック B_s における共分散行列である．学習データが S 個のブロックに分けられ，それぞれのブロックのフレーム数が B である場合，逆行列の計算量は $\mathcal{O}(SB^3)$ となる．

一方で，PIC 近似は局所 GP と同様にブロックごとの共分散行列を用いるが，ブロック対角以外の要素には低ランクの行列による近似を用いる方法である．データ全体の共分散行列は以下のように与えられる．

$$\mathbf{K}_{N+T}^{\mathrm{PIC}} = \mathbf{Q}_{N+T} + \mathrm{blkdiag}\left[\mathbf{K}_{N+T} - \mathbf{Q}_{N+T}\right] \tag{5}$$

行列 $\mathbf{Q}_{N+T}$ は，pseudo-data と呼ばれる M ($\ll N$) 個のデータを用いて以下のように求められる低ランクの行列である．

$$\mathbf{Q}_{N+T} = \mathbf{K}_{(N+T)M}\mathbf{K}_M^{-1}\mathbf{K}_{M(N+T)} \tag{6}$$

ただし行列 $\mathbf{K}_{(N+T)M}$，$\mathbf{K}_{M(N+T)}$ は学習および評価データと pseudo-data との共分散を要素に持ち，行列 $\mathbf{K}_M$ は pseudo-data 内の共分散を要素に持つ．PIC 近似を用いることで，逆行列の計算量を $\mathcal{O}(S(B^3+M^3))$ まで抑えることができ，ブロック間の共分散を使用することができる．

局所 GP および PIC 近似を使用するにはブロックを決定する必要がある．本研究では，HMM 音声認識・合成で広く用いられているコンテキストクラスタリングを用いる．クラスタリングは音素レベルで行い，決定木のリーフノードに含まれる学習データのフレーム数が B を下回ったら，そのノードの分割を停止する．また，本研究では pseudo-data は学習データからランダムに選択する．

4 フレームコンテキストの拡張と畳み込みカーネル

先の報告 [1] では，フレームの入力変数（フレームコンテキスト）に当該音素の音素コンテキストと音素内の相対位置を用いたが，連続音声に対しては不十分である．本稿ではフレームコンテキストおよびカーネル関数を拡張し音素の境界を考慮する．具体的には，先行・当該・後続の音素を基準とした場合のコンテキストの集合として，以下のように定義する．

$$\mathbf{x}_n = \{(w_n^{(i)}, p_n^{(i)}, \mathbf{c}_n^{(i)})|i = -1, 0, 1\} \tag{7}$$

$i = -1, 0, 1$ はそれぞれ先行・当該・後続を表し，$p_n^{(i)}$ は i に対応する音素を基準とした場合の相対位置，$\mathbf{c}_n^{(i)}$ は音素コンテキストである．$w_n^{(i)}$ は重みであり，当該音素に対応する重みを大きくすることによって，当該音素のコンテキストが共分散に対して重要であることを強調できる．

---

*Gaussian-process-regression-based speech synthesis using sparse approximation and convolution kernel, by KORIYAMA, Tomoki, NOSE, Takashi, 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]

Method	Number of training sentences 150	250	350	450
HMM-ML	5.76	5.58	5.46	5.41
HMM-MGE	5.71	5.53	5.41	5.38
GPR-LS	5.49	5.29	5.19	5.13
GPR-PS	5.46	5.27	5.18	5.11
GPR-PM	5.43	5.24	5.15	5.07

拡張したフレームコンテキストの共分散には，2 つの入力に含まれる集合の，すべての組合せの和を計算する畳み込みカーネル [5] を使用する．

$$k(\mathbf{x}_m, \mathbf{x}_n) = \sum_{i\in\{-1,0,+1\}} \sum_{j\in\{-1,0,+1\}} \Big[w_m^{(i)} w_n^{(j)} k_p(p_m^{(i)}, p_n^{(j)}) k_c(\mathbf{c}_m^{(i)}, \mathbf{c}_n^{(j)}) \Big] \quad (8)$$

これによって，音素境界において滑らかに変化する共分散を得ることができる．

5　実験

5.1　実験条件

音声データベースには女性話者一人により発話された ATR 音素バランス文 503 文を用いた．サンプリング周波数 16kHz の音声データに対し，5 ミリ秒毎に STRAIGHT を用いて抽出したスペクトル特徴量をもとに得られた 0 次から 39 次のメルケプストラムを，各フレームの音響特徴量とした．なお，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．データベースの中から 150，250，350，450 文を選んだものをそれぞれ学習データとし，学習データに含まれない 53 文を評価データに用いた．本実験では，メルケプストラムの予測分布の平均系列をメルケプストラムの合成系列とし，音素継続長，F0，非周期性指標には原音声のものを用いた．

カーネルのハイパーパラメータは事前実験により値を決定した．提案法におけるコンテキストクラスタリングには，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) を用いた．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．提案法としては，局所 GP による近似を用いフレームコンテキストを隣接する音素まで拡張しない場合 (GPR-LS)，PIC 近似を使用しフレームコンテキストを拡張しない／した場合 (GPR-PS, GPR-PE) の 3 手法を比較した．従来法には，最尤 (ML) 基準により学習した HMM 音声合成 (HMM-ML) に加え，誤差最小化学習による最適化を行う手法 [6](HMM-MGE) を用いた．

5.2　結果

合成されたメルケプストラムと原音声との平均メルケプストラム距離を表 1 に示す．表から，提案法の 3 手法は従来法の HMM-ML，HMM-MGE に比べ，150 文から 450 文のどの学習データ量においても歪が減少

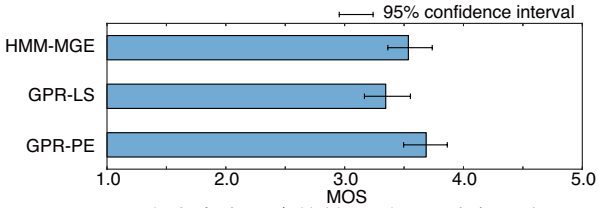

Fig. 1　合成音声の自然性に対する主観評価

していることがわかる．また，GPR-PS は GPR-LS より僅かながら歪が小さく，GPR-PE は GPR-PS より歪が小さくなっている．このことは，PIC 近似および拡張コンテキストが音素間の共分散を表すのに有効に作用したためと考えられる．

次に，MOS により合成音声の自然性を評価した．学習データは 450 文の場合である．被験者は 10 名で，それぞれの被験者は評価データの 53 文からランダムに選択された 10 文に対し，1.bad〜5.excellent の 5 段階で評価を行った．主観評価実験では HMM-MGE，GPR-LS，GPR-PE の 3 手法を比較した．結果を図 1 に示す．GPR-PE は他の手法に比べ僅かではあるが高いスコアを得ることができた．GPR-LS の歪は HMM-MGE より小さかったもののスコアは低くなっている．これは音素間の共分散が無視され，音素境界において生成パラメータが不連続になってしまったためと考えられる．

6　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，発話単位での連続音声の合成を行うための検討を行った．具体的には計算量削減のためにガウス過程の近似法として局所 GP および PIC 近似を導入し，さらに音素境界における共分散を滑らかにするために，フレームコンテキストを拡張し，カーネルに畳み込みカーネルを用いた．スペクトル特徴量の生成実験を行い，HMM に基づく手法に比べ歪が減少することを示した．今後はアクセントなど音素以外のコンテキストの使用や，F0 および音素継続長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，“音声合成のためのガウス過程回帰を用いたフレームレベル音響モデリングの検討，” 音講論 (春)，1-7-5，pp.271-272，2013.

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

[3] H. Wackernagel, *Multivariate Geostatistics*, Springer, 2003.

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

[5] D. Haussler, “Convolution kernels on discrete structures,” Technical Report UCSC-CRL-99-10, Dept of Computer Science, University of California at Santa Cruz., 1999.

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