論文 / 著書情報
Article / Book Information

論題(和文)	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
Title(English)	A study on frame-level acoustic modeling using Gaussian process regression for speech synthesis
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年春季研究発表会講演論文集, Vol. , No. , pp. 271-272
Citation(English)	Proceedings of the ASJ 2013 Spring Meeting, Vol. , No. , pp. 271-272
発行日 / Pub. date	2013, 3

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*

☆郡山知樹，能勢　隆，小林隆夫 (東工大)

1　はじめに

統計的パラメトリック音声合成として知られるHMM 音声合成 [2] は比較的少量のデータベースで利用可能な音声合成手法であり，近年では実用化も進められている．しかし HMM 音声合成ではそのモデル化手法のために合成系列の過剰平滑化などの問題が生じる．一方で，音声のモデル化手法として近年声質変換 [3] や音声表現 [4] にガウス過程 [1] を使用する手法が提案されている．ガウス過程はモデル構造の影響を受けにくいノンパラメトリックベイズモデルとして知られ，ガウス過程によって音声の柔軟なモデル化が可能になるが，テキスト音声合成に対する十分な検討は行われていない．そこで本稿では統計的音声合成の新たな枠組みとして，ガウス過程を用いる手法を提案する．

2　ガウス過程回帰

ガウス過程は回帰分析やクラス分類などの教師あり機械学習に広く使用されているモデルであり，モデルの複雑さに対する柔軟性と過学習に対する頑健性を兼ね備えたノンパラメトリックベイズモデルとして知られている [1]．入力変数 $\mathbf{x}$ に対する出力変数 y がガウス過程に従うとき，正規化された出力データ全体の観測列 $\mathbf{y} = [y_1, \ldots, y_N]^\top$ の確率密度関数は次のガウス分布で表される．

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{0}, \mathbf{K} + \sigma_n^2 \mathbf{I}) \tag{1}$$

ここで，$\mathbf{K}$ は $K_{mn} = k(\mathbf{x}_m, \mathbf{x}_n)$ を要素に持つグラム行列であり $k(\mathbf{x}_m, \mathbf{x}_n)$ は 2 変数間の相関を表すカーネル関数である．また，σ_n は観測データのノイズの大きさを表す．

ガウス過程による回帰分析では未知の入力変数 $\mathbf{x}_*$ に対し連続値の観測変数 y_* の分布を予測する．学習データと未知の入力変数 $\mathbf{x}_*$ の相関を表すグラム行列 $\mathbf{k}_*$ を用いると $\mathbf{y}$ と y_* の同時分布は

$$p\left(\begin{bmatrix}\mathbf{y}\\ y_*\end{bmatrix}\right) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}\mathbf{K} + \sigma_n^2\mathbf{I} & \mathbf{k}_*\\ \mathbf{k}_*^\top & k(\mathbf{x}_*, \mathbf{x}_*) + \sigma_n^2\end{bmatrix}\right) \tag{2}$$

で表される．ゆえに未知の観測変数 y_* の予測分布は以下で与えられる．

$$p(y_*|\mathbf{y}) = \mathcal{N}(\mu_*, \sigma_*^2) \tag{3}$$

$$\mu_* = \mathbf{k}_*^\top (\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} \tag{4}$$

$$\sigma_*^2 = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top (\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{k}_* + \sigma_n^2 \tag{5}$$

ガウス過程による回帰分析のためにはカーネル関数の設計が必要である．カーネル関数に求められる条件はグラム行列が正定値対称行列となることであり，これまでに SE(二乗誤差) カーネルや線形カーネルなど様々なカーネルが提案されている．また，和・積・たたみ込みによって組み合わせた関数もまた，カーネル関数として使用できることが知られている．

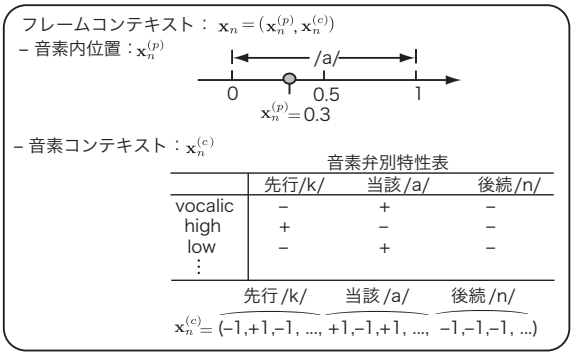

Fig. 1　フレームコンテキスト

Table 1　音素弁別特性

母音性, 高段性, 低段性, 前方性, 後舌性, 舌頂性, 破裂音性, 破擦音性, 継続音性, 有声, 鼻音性, 半母音性, 無音

3　ガウス過程回帰による音声のモデル化

HMM 音声合成における問題点の一つに，状態内で出力確率分布が一定であるという HMM の制約がある．この制約のために短時間に細かく変化する音響特徴量系列を，固定された状態数と動的特徴量で表現することは容易ではない．また，HMM 音声合成では未知のコンテキストに対して頑健なモデルを構築するために木構造に基づくクラスタリングを行うが，クラスタリングにおける平均化処理のために生成されるパラメータ列が過剰に平滑化されるという問題が起きる．

本稿で提案するガウス過程に基づく回帰モデルでは，テキストや書き起こしから得られる各フレームの言語特徴量を入力変数とし，スペクトルなどの各フレームの音響特徴量を出力変数とするガウス過程の枠組みにおいてフレームレベルのカーネル関数を定義することによって，動的特徴量や木構造のクラスタリングを用いずにフレームの音響特徴量を直接モデル化することが可能になる．

3.1　フレームコンテキスト

フレーム間のカーネルを定義するために，まず回帰モデルの入力変数に使用するフレームレベルの言語特徴量を考える．この入力変数をフレームコンテキストと呼び，本稿では初期的な検討事項として図 1 に例を示す単純なフレームコンテキストを使用する．フレームコンテキストは音素内位置と音素コンテキストで構成され，音素内位置は音素の開始フレームを 0，終了フレームを 1 とした相対的な位置で表される．音素コンテキストは先行，当該，後続の音素に対して，表 1 に示す 13 個の音素バランス弁別特徴 [5] の各特徴を $(+1, -1)$ で表した，計 39 次元の二値ベクトルを使用する．

*A study on frame-level acoustic modeling using Gaussian process regression for speech synthesis, by KORIYAMA, Tomoki, NOSE, Takashi, and KOBAYASHI, Takao (Tokyo Institute of Technology)

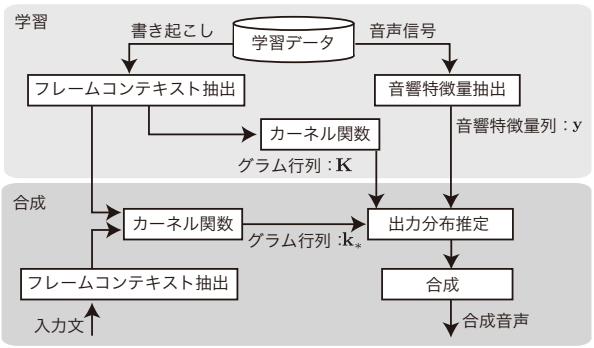

Fig. 2　ガウス過程回帰による音声合成のフロー

3.2　フレームコンテキストカーネル

本稿ではフレームコンテキストの類似度を表すカーネル関数 $k(\mathbf{x}_m, \mathbf{x}_n)$ として位置の類似度を表す位置カーネル $k_p(\mathbf{x}_m^{(p)}, \mathbf{x}_n^{(p)})$ と音素の類似度を表す音素コンテキストカーネル $k_c(\mathbf{x}_m^{(c)}, \mathbf{x}_n^{(c)})$ を掛け合わせたフレームコンテキストカーネルを提案する．

$$k(\mathbf{x}_m, \mathbf{x}_n) = k_p(\mathbf{x}_m^{(p)}, \mathbf{x}_n^{(p)}) k_c(\mathbf{x}_m^{(c)}, \mathbf{x}_n^{(c)}) \quad (6)$$

このカーネルによって，音素内位置および前後の音素情報を考慮した合成を行うことができる．

位置カーネルには音素内位置に対する SE カーネルを使用する．

$$k_p(\mathbf{x}_m^{(p)}, \mathbf{x}_n^{(p)}) = \exp(-((\mathbf{x}_m^{(p)} - \mathbf{x}_n^{(p)})/l_p)^2) \quad (7)$$

ただし l_p はスケールを表すハイパーパラメータである．音素コンテキストカーネルに対しては線形カーネルを使用する．

$$k_c(\mathbf{x}_m^{(c)}, \mathbf{x}_n^{(c)}) = \sum_{i=1}^{3M} \theta_{ci}^2 x_{m,i}^{(c)} x_{n,i}^{(c)} \quad (8)$$

M は音素弁別特性の数であり，θ_{ci} は i 番目の音素特徴の重要度を表すハイパーパラメータを表す．

3.3　音声合成システム

図 2 に音声合成のフローを示す．学習時にはまず音声から音響特徴量を抽出し，対応する書き起こしデータからフレームコンテキストを生成する．そしてフレームコンテキストを用いて学習データのフレーム間のグラム行列 $\mathbf{K}$ を計算する．合成時には入力文から抽出したフレームコンテキストを用いて学習データ・入力データのフレーム間のグラム行列 $\mathbf{k}_*$ を計算する．グラム行列と学習データの観測系列から出力系列の分布を推定し，平均系列 μ_* を合成系列とする．

4　実験

スペクトル特徴量の再現性を客観評価実験により評価した．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 文章を用いた．サンプリング周波数 16kHz で標本化された音声に対し，5 ミリ秒毎に STRAIGHT を用いて抽出したスペクト包絡から得られた 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．ガウス過程回帰に基づく音声合成の基礎的な検討事項として日本語の主要な音素である 5 母音 (/a/, /i/, /u/, /e/, /o/) と 5 つの子音 (/k/, /s/, /t/, /n/, /m/) について音素単位のモデリングを行った．学習用の 450 文章の中から約 10,000 フレームになるように音素セグメントを音素毎にランダムに選択し学習データとし，学習データに含まれない 53 文章から 50 の音素セグメントをランダムに選択しテストデータとした．カーネル関数のハイパーパラメータは事前実験から $\sigma_n = 1.0, l_p = 1.0, \theta_{ci} = 1.0/3M$ $(i = 1, \ldots, 3M)$ とした．

従来手法として通常の HMM 音声合成を比較対象とした．モデルには 5 状態，left-to-right，スキップなしで単一のガウス分布を持つ隠れセミマルコフモデルを用いた．音響特徴量は一次，二次の動的特徴量を含む 120 次元のベクトルとし，図 1 のフレームコンテキストと同等の言語情報を有するトライフォンを用いてクラスタリングを行い，その停止基準には MDL 基準を用いた．

Table 2　合成音声の平均メルケプストラム距離 [dB]

音素	HMM	GPR	音素	HMM	GPR
a	5.67	5.52	k	5.09	5.05
i	6.01	5.63	t	4.13	4.17
u	6.10	5.94	n	5.73	5.81
e	5.33	5.16	s	4.74	4.57
o	5.90	5.64	m	5.48	5.50

合成されたスペクトル列の原音声に対する歪を表 2 に示す．歪には平均メルケプストラム距離を用いた．HMM は HMM 音声合成を GPR はガウス過程回帰を使用した場合をそれぞれ表す．従来手法の HMM と比較すると，母音や音素/s/でガウス過程を用いることで歪が減少していることがわかる．

5　まとめ

本稿ではガウス過程に基づく回帰モデルを用いた音声合成手法を提案した．音素毎の客観評価実験を行い，主に母音に対して HMM に基づく従来手法に比べ高いスペクトル再現性が得られた．今後は実際のテキスト音声合成システムに向け文全体に対するモデル化や韻律情報の導入を行う．

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, 23700195) の助成を得た．

参考文献

[1] 吉村 他，“HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化，” 信学論，J83-D-II(11), 2099–2107, 2000.

[2] N.C.V. Pilkington, et al., “Gaussian process experts for voice conversion,” Proc. INTERSPEECH 2011, 2761–2764.

[3] G.E. Henter, et al., “Gaussian process dynamical models for nonparametric speech representation and synthesis,” Proc. ICASSP 2012, 4505–4508.

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

[5] T. Fukuda and T. Nitta, “Orthogonalized distinctive phonetic feature extraction for noise-robust automatic speech recognition,” IEICE Trans. Inf. & Syst., 87(5), 1110–1118, 2004.