T2R2 東京工業大学リサーチリポジトリ
Tokyo Tech Research Repository

論文 / 著書情報
Article / Book Information

論題	GPR音声合成における深層ガウス過程の利用の検討
Title	On the Use of Deep Gaussian Processes for GPR-based Speech Synthesis
著者	郡山知樹, 小林隆夫
Authors	Tomoki Koriyama, Takao Kobayashi
出典	電子情報通信学会技術研究報告, Vol. 117, No. 517, pp. 27-32
Citation	IEICE Technical Report, Vol. 117, No. 517, pp. 27-32
発行日 / Pub. date	2018, 3
URL	http://search.ieice.org/
権利情報	本著作物の著作権は電子情報通信学会に帰属します。
Copyright	(c) 2018 Institute of Electronics, Information and Communication Engineers

Powered by T2R2 (Tokyo Institute Research Repository)

一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
EA2017-106,SIP2017-115,SP2017-89 (2018-03)

GPR音声合成における深層ガウス過程の利用の検討

郡山　知樹†　　小林　隆夫†

† 東京工業大学工学院　〒226–8502 神奈川県横浜市緑区長津田町 4259 G2–4
E-mail: †{koriyama,takao.kobayashi}@ip.titech.ac.jp

あらまし　本稿では，ベイズ推定の枠組みとして利用されるガウス過程を多層に組み合わせた深層ガウス過程を，統計的音声合成の枠組みに適用する．深層ガウス過程 (DGP) は，ディープニューラルネットワーク (DNN) と同様に深層構造に基づく高精度な予測が期待されるモデルであり，DNN に比べ過学習が起こりにくいという特長を持つ．これまでの報告で，大量のデータに適用可能な二重確率的変分推論 (DSVI) に基づく DGP によって，DNN 音声合成より自然性の高い音声を生成できることが示されたが，モデルの構造などの詳細な検討は行われなかった．本研究では，カーネル関数および層の数，中間層の次元数など様々な条件で音声合成実験を行い，パラメータが音響特徴量歪に与える影響を調査する．

キーワード　深層ガウス過程，統計的パラメトリック音声合成，確率的変分ベイズ

On the Use of Deep Gaussian Processes for GPR-based Speech Synthesis

Tomoki KORIYAMA† and Takao KOBAYASHI†

† School of Engineering, Tokyo Institute of Technology
G2–4, 4259 Nagatsuta-cho, Midori-ku, Yokohama, 226–8502 Japan
E-mail: †{koriyama,takao.kobayashi}@ip.titech.ac.jp

Abstract　This paper proposes a speech synthesis framework based on deep Gaussian processes (DGPs). DGP is a Bayesian deep learning model that is composed of stacked Gaussian process regression. In our preliminary experiments, DGP-based system yielded more natural-sounding synthetic speech than DNN-based one. However, the performance evaluation of DGP had not been done in detail. In this paper, we perform speech synthesis under various experimental conditions with chainging kernel function and the number of layers, and examine the relationships between acoustic feature distortions and model architectures.

Key words　deep Gaussian process, stochastic variational inference, statistical parametric speech synthesis

1. まえがき

近年深層構造を用いた統計モデルは様々な分野に応用され，統計的音声合成の分野においてもディープニューラルネットワーク (DNN) を用いてコンテキストから音響特徴量への写像を学習する手法の有効性が広く知られている [1]．DNN に基づく手法では，パーセプトロンなどの勾配計算の可能な関数を自在に組み合わせることで，多彩なモデルを設計することが可能である．一方でユニット数や層の数，活性化関数など，モデル構造が性能に大きく影響を及ぼすため，DNN の学習には注意深いチューニングが必要という問題がある．

我々はこれまでに，ガウス過程回帰 (GPR) に基づく音声合成 [2] の品質を向上させる手法として，GP-DNN ハイブリッドモデル [3] に基づく手法を提案した [4]．この手法では言語特徴から得られるコンテキストに対して DNN を用いて特徴抽出を行い，その特徴量を GPR の入力とする．GP-DNN ハイブリッドモデルでは，確率的変分推論 (SVI) [5] と呼ばれる近似手法を GPR に用いることで，GPR と DNN のパラメータを同時に最適化することが可能になっている．GPR はベイズ推定のモデルであるため，DNN でしばしば問題となる過学習が起こりにくいモデルと考えられ．実際にこれまでの報告 [4] では GP-DNN ハイブリッドモデルは DNN に比べ自然性の高い音声を合成できることを示した．しかし，GP-DNN ハイブリッドモデルの DNN の部分の重みやバイアスといったパラメータに対して，事前分布を仮定した周辺化を行っているわけではない．そのため，パラメータ数の多い重みやバイアスが学習データに過適応してしまうという DNN の問題は残ったままである．

一方で深層構造に基づくガウス過程の異なる枠組みとして深層ガウス過程 (DGP: deep Gaussian process) [6] が提案されている．DGP は一層でも表現力の高いガウス過程を多層に組

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.
Copyright ©2018　by　IEICE

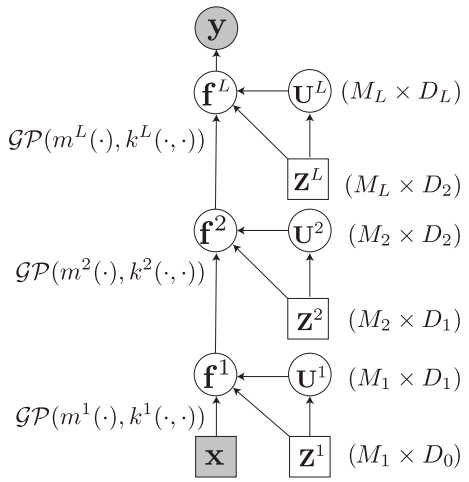


図 1　DSVI-DGP のグラフィカル表現
Fig. 1　Graphical representation of DSVI-DGP.

み合わせたモデルであり，Damianou ら [6] はデータ量が不足している場合でも学習できる深層構造モデルであることを示している．DGP をサイズの大きいデータへの適用する方法は文献 [7] や [8] で提案されており，特に Salimbeni らの手法 [8] では二重確率的変分推論 (DSVI) と呼ばれる近似手法を導入することで，任意のカーネル関数を用いた DGP が大きいサイズの学習データに適用可能であることが示されている．DSVI に基づく DGP では，DNN で用いられる線形変換パラメータの点推定の代わりに，補助変数の変分分布を推定するため，GP-DNN よりさらに過学習が起こりにくいというメリットがある．

文献 [9] で我々は DGP を音声合成に適用し，DNN に対して有意に自然性が高く GP-DNN に対して僅かに自然性が高いという結果を得た．しかし，文献 [9] のシステムでは，従来の GPR 音声合成と同様に，音素や句といった音声単位ごとのコンテキストを使用していたが，DNN 音声合成で広く使用されている，全音声単位のコンテキスト情報を並列に接続したコンテキストとは異なるものだった．そこで本研究では，従来のコンテキストを使用する加算構造のカーネルと，DNN のコンテキストを用いる単一構造のカーネルの比較を行う．また，DGP の学習ではしばしば悪い局所解に陥ってしまい学習が失敗するパターンが存在する．そのため，DGP に用いるカーネル関数，層の数など様々な構造のモデルに対し音声合成実験を行い，どのような場合に学習が失敗しやすいのか，あるいは高精度な推定が行えるのか，といった検討事項に対し，音響特徴量歪の変化を調べる．

2. 深層ガウス過程に基づく音声合成

学習データのデータ点 $i (i = 1 \dots N)$，次元 d の音響特徴量 $\mathbf{y}_i^d$ とフレームレベルのコンテキスト $\mathbf{x}_i$ の関係が，潜在関数 f^d によって $y_i^d = f^d(\mathbf{x}_i) + \epsilon$ と表されるとする．ただし ϵ は分散 σ_ν^2 に従うガウスノイズである．ここで関数 f^d が，平均関数 $m(\cdot)$，カーネル関数 $k(\cdot,\cdot)$ で表されるガウス過程 $\mathcal{GP}(m^d(\mathbf{x}_i), k(\mathbf{x}_i, \mathbf{x}_j))$ に従うことを仮定して，予測を行うのがガウス過程回帰 (GPR) の枠組みである．

深層ガウス過程 (DGP) [6] では，潜在関数が L 層の階層的な合成関数で表されることを仮定し，さらに層 ℓ と $\ell - 1$ との間にガウス過程を仮定する．

$$f^{\ell,d} \sim \mathcal{GP}\left(m^{\ell,d}(\mathbf{f}_i^{\ell-1}), k^\ell(\mathbf{f}_i^{\ell-1}, \mathbf{f}_j^{\ell-1})\right) \tag{1}$$

ただし，$\mathbf{f}_i^{\ell-1}$ は $\ell - 1$ 層までの関数の出力を表す潜在関数変数であり，$D_{\ell-1}$ 次元のベクトルとする．また $\mathbf{f}_i^0 = \mathbf{x}_i$ とする．学習データの全てのフレームの音響特徴量列を，(i, d) 成分に $y_{i,d}$ を持つ行列 $\mathbf{Y}$ と定義し，同様に $\mathbf{F}^l$，$\mathbf{X}$ を定義すると，音響特徴量 $\mathbf{Y}$ の周辺尤度は以下の式で表される．

$$p(\mathbf{Y}|\mathbf{X}) = \int \dots \int p(\mathbf{Y}|\mathbf{F}^L) p(\mathbf{F}^L|\mathbf{F}^{L-1}) \dots p(\mathbf{F}^1|\mathbf{X}) \, d\mathbf{F}^L \dots d\mathbf{F}^1 \tag{2}$$

DGP を音声合成に直接適用することは計算量の問題から困難である．そこで本研究では，大量のデータにも適用可能な手法として提案された DSVI(doubly stochastic variational inference: 二重確率的変分推論) に基づく DGP [8] を用いる．DSVI では，ガウス過程において大量のデータを利用可能な手法として導入された確率的変分近似 [5] を，DGP における複数の層のガウス過程に適用する．DSVI-DGP のグラフィカル表現を図 1 に示す．このモデルでは，それぞれの層 ℓ で $M_\ell (\ll N)$ 個の補助点を使用する．このときガウス過程の仮定から，補助入力 $\mathbf{Z}^\ell = [\mathbf{z}_1^{\ell\top}, \dots, \mathbf{z}_1^{\ell\top}]^\top$ と補助出力 $\mathbf{U} = [\mathbf{u}_1^{\ell\top}, \dots, \mathbf{u}_1^{\ell\top}]^\top = [\mathbf{u}^{\ell,1}, \dots, \mathbf{u}^{\ell,D_\ell}]$ の間には $u_i^{\ell,d} = f^{\ell,d}(\mathbf{z}_i^\ell)$ の関係が与えられる．$\mathbf{u}^{\ell,d}$ の事前分布 $p(\mathbf{u}^{\ell,d}|\mathbf{Z}^\ell)$ に対し，事後分布 $p(\mathbf{u}^{\ell,d}|\mathbf{Z}^\ell, \mathbf{Y}, \mathbf{X})$ を変分分布 $q(\mathbf{u}^{\ell,d}) = \mathcal{N}(\mathbf{u}^{\ell,d}; \mathbf{m}^{\ell,d}, \mathbf{S}^{\ell,d})$ で近似すると，周辺尤度の変分下限は以下のようになる．

$$\mathcal{L} = \sum_{i=1}^{N} \left\{ \sum_{d=1}^{D_L} \mathbb{E}_{q(f_i^{L,d})} \left[\log p(y_i^{L,d}|f_i^{L,d}) \right] - \frac{1}{N} \sum_{\ell=1}^{L} \sum_{d=1}^{D_\ell} \mathrm{KL}(q(\mathbf{u}^{\ell,d}) \| p(\mathbf{u}^{\ell,d}|\mathbf{Z}^\ell)) \right\} \tag{3}$$

DSVI-DGP ではこの変分下限を最大化することでモデルの学習を行う．式 (3) 右辺の第一項はモデルの適合度を表し，第二項は補助点の過学習を避けるペナルティ項である．このとき変分下限がサンプル点ごとの値の和で表されるため，確率的勾配上昇法に基づく最適化が可能である．

ただし，式 (3) の変分事後分布 $q(\mathbf{f}_i^L) = \prod_{d=1}^{D} q(f_i^{L,d})$ は

$$\begin{aligned} q(\mathbf{f}_i^L) &= \int \prod_{\ell=1}^{L} \int p(\mathbf{f}_i^\ell|\mathbf{f}_i^{\ell-1}, \mathbf{U}^\ell) q(\mathbf{U}^\ell) d\mathbf{U}^\ell d\mathbf{f}_i^\ell \\ &= \int \prod_{\ell=1}^{L} q(\mathbf{f}_i^\ell|\mathbf{f}_i^{\ell-1}) d\mathbf{f}_i^\ell \end{aligned} \tag{4}$$

となり，この積分は特定のカーネル関数を除いて解析的に計算不可能である．そこで DSVI-DGP では，変分オートエンコーダの枠組みで用いられる reparameterization trick [10] と同様に，サンプリングによって下限の近似を行う．具体的には，下

位層のサンプル点 $\hat{\mathbf{f}}_i^{\ell-1}$ が与えられたときの上位層の潜在関数変数 $f_i^{\ell,d}$ の分布 $q(f_i^{\ell,d}|\hat{\mathbf{f}}_i^{\ell-1})$ はガウス分布で表されるため，その平均と分散からサンプリングを行う．S 個のサンプル点を用いると変分事後分布 $q(\mathbf{f}_i^L)$ は，以下の混合ガウス分布で近似される．

$$q(\mathbf{f}_i^L) = \frac{1}{S}\sum_{s=1}^{S} q(\mathbf{f}_i^L|\hat{\mathbf{f}}_{i,s}^{L-1}) \tag{5}$$

これを式 (3) に代入することで変分下限を近似する．

以上から，DSVI-DGP で最適化を行うパラメータは各層の補助入力 $\mathbf{Z}^\ell$，補助出力の変分分布パラメータ $(\mathbf{m}^{\ell,d}, \mathbf{S}^{\ell,d})$，各層のカーネル関数のパラメータ，およびノイズのパワーを表すパラメータ σ_ν^2 となる．一方で，層の数 L や各層のカーネル関数や，補助点数 M_ℓ，各層の次元 D_ℓ はモデル構造を決定するメタパラメータとなる．

音声合成時には，予測変分分布 $q(\mathbf{f}_i^L)$ から音響特徴量 $\mathbf{y}_i$ の予測分布を求めるが，学習時と同様に $q(\mathbf{f}_i^L)$ を解析的に求めることができない．文献 [8] ではサンプリング結果をそのまま出力としているが，本研究では初期的な検討として，予測平均をそのまま使用する．具体的には，各層の予測平均 $\mathbb{E}[q(\mathbf{f}_i^\ell|\hat{\mathbf{f}}_i^{\ell-1})]$ をサンプリング結果と見なして，$L-1$ 層までの値 $\hat{\mathbf{f}}_i^{\ell-1}$ を計算する．その後 1 層の GPR と同様に音響特徴量の予測分布を求め，その分布から音声パラメータ生成を行う．

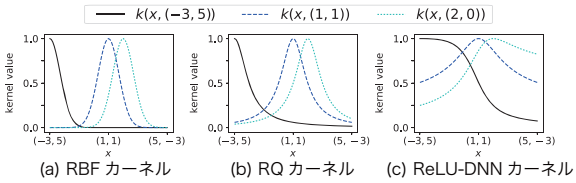
3. 深層ガウス過程における検討事項

文献 [11] で示されるように，ガウス過程は一次元のシンプルなモデルであっても，周辺尤度は多峰になる性質を持っている．このとき，周辺尤度の局所解には，学習データの出力がすべてノイズと判断され予測分布の平均がほとんどすべての点で 0 となってしまうパラメータが存在する．この局所解は DSVI-DGP においても存在すると考えられる．例えば，$L-1$ 層でサンプリングされた変数 $\hat{\mathbf{f}}^{L-1}$ が音響特徴量 $\mathbf{y}$ を適切に表現できない場合，ノイズのパワーが大きくなるように更新され，一方でカーネル関数のパラメータや補助入力が適切に更新されないという現象が起きる．この現象が学習の初期段階で発生すると「悪い局所解」に収束してしまう．特に層が深い場合にはサンプリングによるばらつきが大きくなるため，このような問題が起きる可能性が高いと考えられる．そのため本研究ではモデル構造やカーネル関数を工夫することによって，悪い局所解を避けることを検討する．

3.1 カーネル構造

本研究では，最上位層のカーネル構造として，単一構造のカーネルと加算構造のカーネルを検討する．単一構造のカーネルはフィードフォワード型の DNN 音声合成と同様に，フレームレベルの情報を全て接続したコンテキストを単一の DGP でモデル化する．単一構造のモデルはシンプルであるが，コンテキストが多様な情報を含む高次情報となるため，DGP の低層におけるコンテキストの特徴抽出が初期の段階で正常に行われず，前節で述べた悪い極所解に陥る可能性がある．

一方で加算構造のカーネルは，これまでの GPR 音声合成 [2] および GP-DNN ハイブリッドモデルに基づく音声合成 [4] で使用されたカーネルである．加算構造のカーネルでは，「音素の開始」や「アクセント句の終了」など時系列上の瞬時的なイベントに基づくコンテキストに対し，イベントごとに DGP でコンテキストの特徴抽出を行う．最上位層ではイベントごとの部分カーネルの重み付き和によるカーネルを用いる．このとき部分カーネルの入力変数には DGP の出力変数を使用する．



(a) RBF カーネル　(b) RQ カーネル　(c) ReLU-DNN カーネル

図 2　カーネル関数の例
Fig. 2　Example of kernel functions.

加算構造のカーネルでは，音声単位ごとにコンテキストの特徴抽出を行うため，単一構造のカーネルに比べモデルへの依存度が低い．しかし，単一構造のカーネルに比べ音声単位間の関係のモデル化が困難である．

3.2 カーネル基本関数

本研究では，カーネルに用いる基本関数として，広く用いられている RBF カーネルだけでなく，RQ カーネル，ReLU-DNN カーネルの検討を行う．具体的には，単一構造の各層および加算構造の中間層で用いるカーネルと，加算構造の最上位層における部分カーネルに用いるカーネル関数を検討する．また，DGP において最上位層は音響特徴量の推定，下位層はコンテキストの特徴抽出の機能を持つと見なせるため，最上位層と下位層で異なるカーネルを用いることを考える．

それぞれのカーネルに対し，2 次元の入力変数空間において $x_1 + x_2 = 2$ の直線上のカーネル関数の例を図 2 に示す．

3.2.1 RBF カーネル

RBF カーネルはベクトル間の距離に基づいて決定される代表的な stationary（不変）カーネルで，SE (squared exponential) カーネルやガウスカーネルとも呼ばれる．RBF カーネルは D 次元の入力 $\mathbf{x}, \mathbf{x}'$ に対して以下の式で定義される．

$$k_{\mathrm{RBF}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\Delta(\mathbf{x}, \mathbf{x}')}{2}\right) \tag{6}$$

$$\Delta(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')^\top \mathbf{L}^{-1} (\mathbf{x} - \mathbf{x}') \tag{7}$$

$$\mathbf{L} = \mathrm{diag}[l_1^2, \ldots, l_D^2] \tag{8}$$

l_d はスケールパラメータであり，次元 d の重要度を決定する．RBF カーネルの問題点の一つは 2 つのベクトル間の距離が大きいとき，値が 0 に近い値になってしまうことである．そのため，学習時に 2 点間の距離が遠すぎる場合に適切なパラメータ更新が行われず，ノイズのパワーだけが大きくなってしまう現象が起こる可能性がある．

3.2.2 RQ カーネル

RQ (rational quadratic) カーネルはスケールの異なる無限個の RBF カーネルの和の形で定義され，以下の式で表される．

$$k_{\mathrm{RQ}}(\mathbf{x},\mathbf{x}') = \left(1 + \frac{\Delta(\mathbf{x},\mathbf{x}')}{2\alpha}\right)^{-\alpha} \tag{9}$$

$\alpha(>0)$ は関数の形状を表すパラメータで $\alpha \to \infty$ とすると RBF カーネルと一致する．RQ カーネルは距離が大きい場合でも 0 に近づく速度が遅いため，パラメータが更新されない現象が起こりにくいと考えられる．

3.2.3 ReLU-DNN カーネル

素子数を無限個にした DNN はガウス過程に一致することが報告されており [12]，この時に得られるカーネル関数は，サポートベクタマシンや GPR においてその有効性が報告されている [12, 13]．ReLU-DNN カーネル [13] は無限個の素子と ReLU 活性化関数のニューラルネットワークから得られるカーネルであり，以下の漸化式で表される．

$$k_0(\mathbf{x},\mathbf{x}') = \sigma_{b0}^2 + \sigma_{w0}^2 \mathbf{x}^\top \mathbf{x}' \tag{10}$$

$i = 1 \dots P$ に対し

$$k_i(\mathbf{x},\mathbf{x}') = \sigma_{bi}^2 + \sigma_{wi}^2 \sqrt{k_{i-1}(\mathbf{x},\mathbf{x})}\sqrt{k_{i-1}(\mathbf{x}',\mathbf{x}')} \cdot (\sin\theta_{i-1} + (\pi - \theta_{i-1})\cos\theta_{i-1})) \tag{11}$$

$$\theta_{i-1} = \cos^{-1}\frac{k_{i-1}(\mathbf{x},\mathbf{x}')}{\sqrt{k_{i-1}(\mathbf{x},\mathbf{x})}\sqrt{k_{i-1}(\mathbf{x}',\mathbf{x}')}} \tag{12}$$

$$k_{\mathrm{ReLUDNN}} = \frac{k_P(\mathbf{x},\mathbf{x}')}{\sqrt{k_P(\mathbf{x},\mathbf{x})}\sqrt{k_P(\mathbf{x}',\mathbf{x}')}} \tag{13}$$

ただし，P は DNN の層数である．本研究では式 (13) で示すカーネルの正規化を最終層の出力に対して行う．ReLU-DNN カーネルは 2 点の位置に基づく nonstationary カーネルであり，式 (11) からもわかるように 2 点を示すベクトルの長さが等しいとき，ベクトル間の角度が小さいほど値が大きくなる．ReLU-DNN カーネルは図 2(c) に示すように RBF カーネルや RQ カーネルとは異なる性質を持っており，シグモイド関数のような形状を持つこともある．

4. 実　　験

4.1 実験条件

データベースには音声合成システム XIMERA [14] に含まれる女性話者 F009 を使用した．学習データには 1593 文 (約 119 分)，評価データには 60 文 (約 4.1 分) の音声を用いた．この中には音素バランス文に加え，旅行会話文，新聞読み上げ文が含まれている．サンプリングレート 16kHz の音声信号から，5ms ごとに STRAIGHT を用いて F0，スペクトル包絡，非周期性指標を抽出し，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，およびそれらの Δ, Δ^2 と有声／無声フラグで構成される 139 次元ベクトルを音響特徴量として使用した．音素継続長を予測する音素単位のモデルと音響特徴量を予測するフレーム単位のモデルを学習した．すべての入力変数および連続値の出力変数に対して平均 0，分散 1 に正規化を行った上でモデルの学習を行った．単一構造のカーネルの入力特徴量は 574 次元とし，加算構造のカーネルでは音素，モーラ，アクセント句，呼気段落および発話の入力変数の次元はそれぞれ 249，88，142，82，41 とした．合成時には，動的特徴量からのパラメータ生成 [15] を行った．DGP の最上位層の補助点数 M_L は 1024 とした．客観評価の指標にはメルケプストラム距離 (MCEP)[dB]，対数 f_0 の RMSE(F0)[cent]，有声／無声誤り率 (V/UV)[%]，非周期性指標歪 (BAP)[dB]，音素継続長の RMSE(DUR)[ms] を用いた．

表 1　DGP および GP-DNN における合成音声の原音声に対する音響特徴量歪

Table 1　The acoustic feature distortions between original and synthetic speech using DGP and GP-DNN.

手法	構造	MCEP	F0	V/UV	BAP	DUR
GP-DNN	単一	4.88	167	4.58	3.16	F
DGP	単一	4.75	159	4.42	3.10	F
GP-DNN	加算	4.89	180	4.55	3.22	16.8
DGP	加算	4.82	176	4.46	3.18	15.4

カーネル基本関数のパラメータの初期値として，RBF カーネルおよび RQ カーネルのスケールは $l_1 = \cdots = l_D = 2$，RQ カーネルにおける α は 1 とした．また ReLU-DNN カーネルの層数 P は 3 とし，$\sigma_{b0} = \cdots = \sigma_{b3} = 1$，$\sigma_{w0} = \cdots = \sigma_{w3} = 1$ とした．また学習時のパラメータ最適化には ADAM [16] を使用し，学習係数は 0.01 とした．DGP は過学習の起こりにくい手法であるため学習の終了条件の決定には検討の余地があるが，本研究では事前実験においてある程度の収束が見られたエポック数を終了条件として用いた．具体的には，音素単位のモデルでは 300 エポック，フレーム単位のモデルでは 30 エポックとした．

DGP の平均関数 $m^\ell(\cdot)$ は先行研究 [8] と同様に，$\ell = 1$ では PCA に基づく次元削減のマッピング関数，$\ell = 2 \dots L-1$ では $m^\ell(\mathbf{x}) = \mathbf{x}$，$\ell = L$ では $m^\ell(\mathbf{x}) = \mathbf{0}$ とした．補助入力 $\mathbf{Z}^\ell$ は，それぞれの層における K-means クラスタリングに基づくセントロイドを初期値として用いた．また．補助出力の変分分布のパラメータの初期値は $\mathbf{m}^{\ell,d} = \mathbf{0}$，$\mathbf{S}^{\ell,d} = \mathbf{I}$ とし，最上位層以外では $\mathbf{S}^{\ell,d}$ は対角行列とした．

4.2 GP-DNN およびカーネル構造の比較

本研究ではまず，DGP と同様にガウス過程に深層構造を導入したモデルである GP-DNN ハイブリッドモデル [4] との比較を行った．GP-DNN における DNN の隠れ層は 256 次元とし，出力層は 32 次元とした．また勾配消失問題の起こりにくい手法として提案されている SELU [17] を活性化関数として使用した．DGP および GP-DNN のカーネル基本関数には RBF カーネルを使用し，DGP は 6 層，GP-DNN は隠れ層 4 層と線形変換層 1 層，GP 層 1 層の計 6 層の構造を用いた．また，DGP の中間層の補助点数は 256，次元は 32 とした．

結果を表 1 に示す．表中の F は 3 節で述べた悪い局所解に陥りすべての出力が 0 となったことを表す．GP-DNN と DGP を同じカーネルの構造同士で比較すると，音素継続長を除くすべての特徴量で DGP の歪が小さいことがわかる．また，カーネルの構造を比較すると，DGP では単一構造のカーネルを用いることで加算構造のカーネルに比べ対数 f_0 の RMSE が 17cent

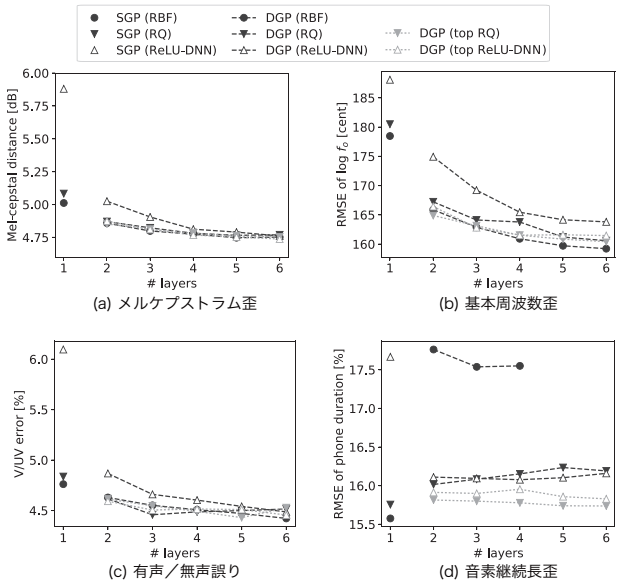


図 3　DGP の層数と音響特徴量歪の関係

Fig. 3　Acoustic feature distortions as a function of the number of DGP layers.

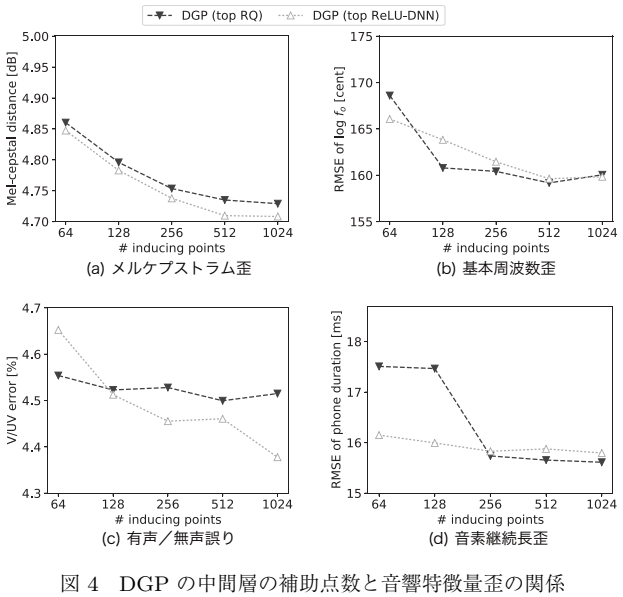


図 4　DGP の中間層の補助点数と音響特徴量歪の関係

Fig. 4　Acoustic feature distortions as a function of the number of inducing points.

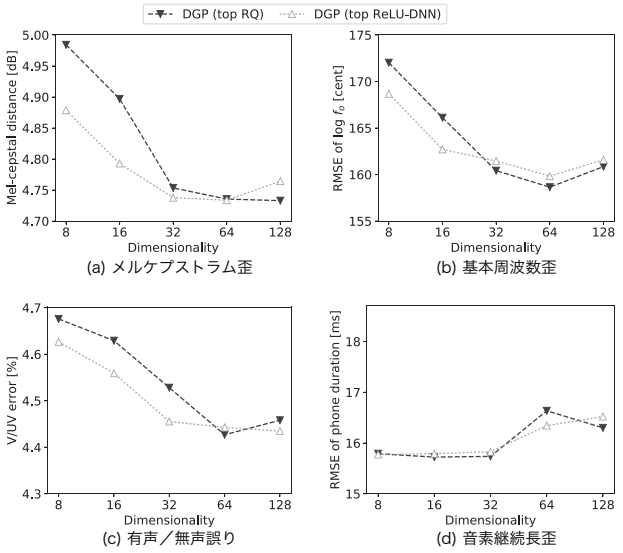


図 5　中間層の次元と音響特徴量歪の関係

Fig. 5　Acoustic feature distortions as a function of the dimensionality of hidden layers.

程度減少するなど，単一構造のカーネルの方が歪が小さくなる傾向が見られた．これは加算構造のカーネルでは音素とアクセント句など音声単位間の関係をモデル化することが難しかったためと考えられる．

本稿ではこの後，スペクトル歪および f_o 歪の最も小さかった，単一構造のカーネルを用いる DGP に注目し検討を行う．

4.3　カーネル基本関数および層の数

DGP の特徴について詳細な検討を行うため，カーネル基本関数には 3.2 節で述べた，RQ カーネルと ReLU-DNN カーネルを RBF カーネルの代わりに使用した．また，DGP の層数を 2〜6 に変化させた．このとき DGP の中間層の補助点数は 256，次元は 32 に固定した．

図 3 に各手法における音響特徴量歪を示す．なお紙面の都合上省略するが，非周期性指標歪の変化の傾向はメルケプストラム歪と非常に似ていた．図中の SGP は $L = 0$ の single-layer GP である．DGP(RBF) と DGP(RQ)，DGP(ReLU-DNN) ではすべての層にそれぞれのカーネルを使用し，DGP(top RQ) と DGP(top ReLU-DNN) では最上位層のみにそれぞれ RQ カーネルと ReLU-DNN カーネルを用いて，それ以外の層は RBF カーネルとした．

図 3(e) に示す音素継続長の予測結果を見ると，DGP(RBF) が 4 層以下の場合にはモデルパラメータが悪い局所解に陥らなかった．しかし，DGP(RBF) の 2〜4 層の場合は他のカーネルに比べ音素継続長歪が大きく，最上位層に RBF カーネルを用いると学習が困難になる可能性があることが示された．また，最上位層以外のカーネル基本関数を比較すると，DGP(ReLU-DNN) は DGP(top ReLU-DNN) に比べ歪が大きくなる場合が多く，中間層のカーネル関数は RBF で十分であると考えられる．

層の数を比較すると，音素継続長歪を除いて，2 層の DGP は SGP より歪が小さく，さらに隠れ層を増やすことで歪が減少していく傾向があることがわかった．

4.4　補助点数と中間層の次元

次に DGP のメタパラメータである中間層の補助点数 $M_\ell\,\ell = (1 \ldots L-1)$ と次元 $D_\ell\,\ell = (1 \ldots L-1)$ の影響を調査した．次元が大きいほど中間層の表現力が向上するが，次元の呪いの問題が起こりやすい．補助点数は代表点の数を表すため，一般的に補助点が多いほど精度が上がるが，学習時の計算量のオーダーは補助点数の 3 乗に比例するというトレードオフがある．

カーネル関数の構成は前節で歪の小さかった DGP(top RQ) と DGP(top ReLU-DNN) を使用し，層の数は 6 とした．中間層の次元を 32 に固定したときの，補助点数と音響特徴量歪の関係を図 4 に示す．図より，どの特徴量においても補助点数が

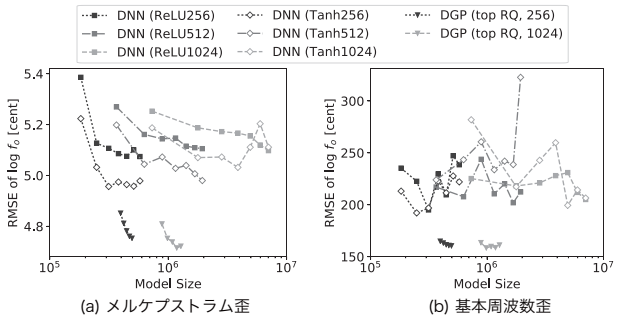


図 6　DGP と DNN における音響特徴量歪の比較

Fig. 6　Comparison of DGP and DNN in terms of acoustic feature distortions.

増加することによって歪が減少する傾向があることが確認できる．また，中間層の補助点数を 256 に固定したときの次元と音響特徴量歪の関係を図 5 に示す．図に示すように，フレームレベルの音響特徴量では次元が多い方が歪が小さくなる傾向があるが，音素継続長歪は次元が少ない方が小さいという結果となった．このような結果となった理由として，音素レベルのモデルでは出力変数が音素継続長の 1 次元であるのに対し，フレームレベルのモデルでは 139 次元であることが考えられる．

4.5　DGP と DNN の比較

最後に DGP および DNN に基づく音声合成の比較を行った．DGP では，DGP(top RQ) のカーネルを使用し，中間層の補助点数に 256 と 1024 を用いた．また層の数を 2 から 6 まで変化させた．DNN には，活性化関数に ReLU と tanh を使用し，ユニット数 256，512，1024 に対して，隠れ層の数を 1 から 7 まで変化させた全 42 パターンを用いた．

図 6 に DGP と DNN のメルケプストラム歪，基本周波数歪を示す．メルケプストラム歪の結果から，2 層の DGP であっても DNN のどの組合せよりも歪が小さいことがわかる．また，基本周波数歪の結果では，DNN では隠れ層の数が増加したときに歪が大きく上昇してしまうといった現象が見られるが，DGP ではそのようなことは見られず，歪のばらつきも DNN に比べ小さいものとなっている．

5.　む　す　び

本研究では DGP を統計的パラメトリック音声合成へ適用する際の知見を得るために，様々なモデル構成で実験を行い，音響特徴量歪との関係を調査した．実験から，複雑な加算構造のカーネルよりも単純な単一構造のカーネルが有効であることを示した．また最上位層のカーネルに RBF カーネルではなくカーネルの値が 0 になりにくい RQ カーネルや ReLU-DNN カーネルを用いることで悪い局所解に陥る問題が見られなくなった．さらに，隠れ層の数や中間層の補助点数を増やすことで歪が減少する傾向にあることを示した．DNN との比較では，DNN の様々なモデル構造より歪が小さいことを示した．一方で，RQ カーネルや ReLU-DNN カーネルの比較など歪だけでは優劣の判別が困難な違いもあった．

実験全体を通して DGP の手法間の歪の差異は小さいため，頑健性の高い学習が可能であることが予想される．そのため，データサイズや音声データの種類によるモデル構造のチューニングの必要性などの検討や，聴覚上での合成音声の評価が今後の課題として挙げられる．

6.　謝　　辞

本研究は JSPS 科研費 JP15H02724, JP17K12711 の助成を受けた．

文　　献

[1] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” Proc. ICASSP, pp.7962–7966, 2013.

[2] T. Koriyama, T. Nose, and T. Kobayashi, “Statistical parametric speech synthesis based on Gaussian process regression,” IEEE Journal of Selected Topics in Signal Processing, vol.8, no.2, pp.173–183, 2014.

[3] J. Bradshaw, A.G.d.G. Matthews, and Z. Ghahramani, “Adversarial examples, uncertainty, and transfer testing robustness in Gaussian process hybrid deep networks,” Proc. Reliable Machine Learning in the Wild - ICML 2017 Workshop, 2017.

[4] 郡山知樹, 小林隆夫, “GP-DNN ハイブリッドモデルに基づく統計的音声合成の検討,” 信学技報, SP2017-67, 2018.

[5] M.D. Hoffman, D.M. Blei, C. Wang, and J. Paisley, “Stochastic variational inference,” The Journal of Machine Learning Research, vol.14, no.1, pp.1303–1347, 2013.

[6] A. Damianou and N. Lawrence, “Deep Gaussian processes,” Proc. AISTATS, pp.207–215, 2013.

[7] Z. Dai, A. Damianou, J. González, and N. Lawrence, “Variational auto-encoded deep Gaussians processes,” Proc. ICLR, 2016.

[8] H. Salimbeni and M. Deisenroth, “Doubly stochastic variational inference for deep Gaussian processes,” Proc. NIPS, pp.4591–4602, 2017.

[9] 郡山知樹, 小林隆夫, “GPR 音声合成における深層構造の利用の検討,” 音講論 (春), 3-8-6, 2018.

[10] D.P. Kingma and M. Welling, “Auto-encoding variational Bayes,” Proc. ICLR, 2014.

[11] C.E. Rasmussen and C.K.I. Williams, Gaussian Processes for Machine Learning, The MIT press, 2006.

[12] J. Lee, Y. Bahri, R. Novak, S.S. Schoenholz, J. Pennington, and J. Sohl-Dickstein, “Deep neural networks as Gaussian processes,” Proc. NIPS 2017 Workshop Bayesian Deep Learning, 2017.

[13] Y. Cho and L.K. Saul, “Kernel methods for deep learning,” Proc. NIPS, pp.342–350, 2009.

[14] 河井恒, 戸田智基, 山岸順一, 平井俊男, 倪晋富, 西澤信行, 津崎実, 徳田恵一, “大規模コーパスを用いた音声合成システム XIMERA,” 電子情報通信学会論文誌 D, vol.89, no.12, pp.2688–2968, 2006.

[15] T. Koriyama, T. Nose, and T. Kobayashi, “Parametric speech synthesis based on Gaussian process regression using global variance and hyperparameter optimization,” Proc. ICASSP, pp.3862–3866, 2014.

[16] D.P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” Proc. ICLR, 2015.

[17] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, “Self-normalizing neural networks,” Proc. NIPS, pp.972–981, 2017.