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	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
Title	A study on statistical speech synthesis based on GP-DNN hybrid model
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. 393, pp. 5-10
Citation	, Vol. 117, No. 393, pp. 5-10
発行日 / Pub. date	2018, 1
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 (Science Tokyo 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
SP2017-67 (2018-01)

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

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，ガウス過程回帰 (GPR) に基づく音声合成の新しいアプローチを提案する．従来の GPR に基づく音声合成は，近似のために木構造によるブロック分割を用いていることから，性能が木構造による予測に依存するという問題があった．そこで本研究では，確率的勾配降下法により効率的な学習の可能な確率的変分ガウス過程 (SVGP) と，コンテキストの特徴抽出器としてのディープニューラルネットワーク (DNN) を組み合わせたハイブリッド手法を提案する．客観評価と主観評価の実験結果から、提案手法はブロック分割を用いた従来の GPR 音声合成や DNN に基づく音声合成に比べ自然な音声が合成可能であることを示す．

キーワード　ガウス過程回帰，確率的変分ベイズ，ニューラルネットワーク，統計的パラメトリック音声合成

A Study on Statistical Speech Synthesis Based on GP-DNN Hybrid Model

Tomoki KORIYAMA† and Takao KOBAYASHI†

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

Abstract　We propose a novel approach to Gaussian process regression (GPR)-based speech synthesis in this paper. Since the conventional GPR-based speech synthesis was based on data partition with a decision tree, a decision tree was bottleneck of the performance of synthetic speech. In contrast, we propose a hybrid model of Gaussian process and deep neural network (DNN). In the hybrid model, DNN extracts context-derived features and the output of DNN is used as an input of Gaussian process. The parameters of DNN and GP are optimized using a minibatch-based stochastic gradient descent method. From the subjective evaluation results, it can be seen that the proposed technique outperforms not only the conventional GPR-based speech synthesis with decision trees but also DNN-based speech synthesis.

Key words　Gaussian process regression, stochastic variational inference, neural network, statistical parametric speech synthesis

1. まえがき

近年ディープニューラルネットワーク (DNN) やリカレントニューラルネットワーク (RNN) に基づく統計的音声合成が研究の主流となっている [1,2]．DNN の利点の一つは，音響特徴量と言語特徴量の複雑な関係を深層構造によって自動的にモデル化できるという点である．また，DNN ではミニバッチに基づく確率的勾配降下法によって，学習データ量に対し線形の計算量で効率的にモデルを学習できる．しかし，DNN はモデル構造やメタパラメータの選択に性能が依存し，過学習がしばしば発生するなどの問題がある．

これに対し，我々は統計的音声合成へのアプローチとして，ガウス過程回帰 (GPR: Gaussian process regression) に基づく音声合成を提案している [3,4]．GPR はノンパラメトリックベイズモデルであり，モデルの柔軟性と過学習に対する頑健性を持ったモデルとして知られている．しかし，GPR の計算量は学習データ量に対して 3 乗のオーダーとなるため，実用上は何らかの近似手法が必要となる．これまでの研究 [3] では，フレーム間の相関関係を表すグラム行列をブロック対角行列と低ランク行列の和で近似する PIC(partially independent conditional) 近似を用いる手法を提案した．PIC 近似では GPR のノンパラメトリック性を活用できるが，ブロック対角のために学習データを分割する方法に性能が依存するという問題点がある．

一方，PIC 近似以外のガウス過程の近似手法として，確率的変分ガウス過程 (SVGP: stochastic variational Gaussian process) が提案されている [5,6]．SVGP では学習データの周辺尤

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

度をサンプル点ごとの和で表される変分下限に変分近似する．このモデルでは確率的勾配降下法によるパラメータ最適化が可能であるため，DNN と同様に大量の学習データに対しても効率的な学習ができる．また，SVGP はベイズ推定の枠組みであるため，過学習が起こりにくいという特長がある．さらに，SVGP を拡張した手法として，近年 SVGP と深層構造を組み合わせた手法が提案されている [7–9]．これらの手法の主な目的は，ガウス過程回帰システム構築時に問題となるカーネル設計の代わりに，深層構造を用いて適切なカーネル関数の入力を抽出することである．

本研究では文献 [8] で提案された SVGP と DNN のハイブリッド手法である GP-DNN に注目し，GP-DNN に基づく GPR 音声合成を提案する．提案手法では，従来法と同様に音素弁別特性などの特徴量やフレームの相対位置をコンテキストとして，コンテキストの類似度に基づくカーネル関数をガウス過程の共分散関数として用いる．このとき，コンテキストを直接カーネル関数への入力に使用せず，コンテキストを DNN によって変換して得られた特徴ベクトルをカーネル関数への入力に使用する．

本稿では，まず確率的変分ガウス過程 (SVGP) について説明し，SVGP を音声合成への適用について述べる．その後，SVGP に基づく音声合成を GP-DNN ハイブリッドモデルに基づく音声合成へと発展させる．客観及び主観評価実験においては，SVGP や従来法の PIC 近似だけでなく DNN および RNN 音声合成との比較を行う．

2. 確率的変分ガウス過程

本節ではガウス過程の近似手法の一つである確率的変分ガウス過程 (SVGP [(注1)]) について述べる．学習データの出力変数および入力変数をそれぞれ $\mathbf{y} = [y_1, \ldots, y_N]^\top$，$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ とする．パラメトリック音声合成の枠組みでは y_i はフレームレベルの音響特徴量，$\mathbf{x}_i$ はフレームレベルのコンテキストを表す．潜在関数 f が $\mathcal{GP}(0, \kappa(\mathbf{x}, \mathbf{x}'))$ で表されるガウス過程に従うとき，出力変数 $\mathbf{y}$ の周辺尤度は

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})d\mathbf{f} \tag{1}$$

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K}_N) \tag{2}$$

$$\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)] \tag{3}$$

で表される．このとき，$\mathbf{K}_N$ はカーネル $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ を (i, j) 要素にもつグラム行列である．ここで出力変数 y_i が以下のように要素毎の条件付き独立であると仮定する．

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{N} p(y_i|f(\mathbf{x}_i)) \tag{4}$$

ここで，ガウス過程のスパース近似でしばしば用いられる，M 点 ($M \ll N$) の疑似データセットの入力変数 $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_M)$（補助点と呼ばれる）および出力変数 $\mathbf{u} = [u_1, \ldots, u_M]^\top$ を導入する．ガウス過程はいかなる入力に対しても出力変数の同時分布がガウス分布になることを保証するため，学習データと疑似データの同時分布は次のに示すガウス分布で与えられる．

$$p(\mathbf{f}, \mathbf{u}) = \mathcal{N}\left(\begin{bmatrix}\mathbf{f}\\ \mathbf{u}\end{bmatrix}; \mathbf{0}, \begin{bmatrix}\mathbf{K}_N & \mathbf{K}_{NM}\\ \mathbf{K}_{MN} & \mathbf{K}_M\end{bmatrix}\right) \tag{5}$$

伝統的な手法である DTC 近似 [11] や FITC 近似 [12]，これまでの GPR 音声合成 [3] で用いている PIC 近似 [13] などの手法では，グラム行列を Nyström 近似により低ランク化することで計算量の削減を行っていた．しかし，これらの手法では目的関数が周辺尤度関数と異なってしまい，結果として補助点に対して過学習する可能性がある．

一方で SVGP では周辺尤度関数自体に対して近似を行う．具体的には変分ベイズの枠組みにおいて，補助変数の事後分布 $p(\mathbf{u}|\mathbf{y})$ を変分分布 $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}; \mathbf{m}, \mathbf{S})$ で近似する．このとき対数周辺尤度の変分下限は以下の式で与えられる．

$$\begin{aligned}\log p(\mathbf{y}) &= \log \iint p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})d\mathbf{f}d\mathbf{u}\\ &\geqq \int q(\mathbf{u}) \left\{\log \frac{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})d\mathbf{f}}{q(\mathbf{u})}\right\} d\mathbf{u}\\ &= \int q(\mathbf{u}) \left\{\log \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})d\mathbf{f}\right\} d\mathbf{u}\\ &\quad - \mathrm{KL}(q(\mathbf{u})\|p(\mathbf{u}))\end{aligned} \tag{6}$$

ここでさらに，イェンセンの不等式を用いることで以下の変分下限 $\mathcal{L}$ を得る．

$$\begin{aligned}\log p(\mathbf{y}) &\geqq \iint q(\mathbf{u})p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f})d\mathbf{f}d\mathbf{u} - \mathrm{KL}(q(\mathbf{u})\|p(\mathbf{u}))\\ &= \int q(\mathbf{f}) \log p(\mathbf{y}|\mathbf{f})d\mathbf{f} - \mathrm{KL}(q(\mathbf{u})\|p(\mathbf{u})) \triangleq \mathcal{L}\end{aligned} \tag{7}$$

$q(\mathbf{f})$ は潜在変数に対する変分分布であり，$q(\mathbf{f}) \triangleq p(\mathbf{f}|\mathbf{u})q(\mathbf{u})$ とすると，以下のガウス分布で表される．

$$q(\mathbf{f}) = \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{8}$$

$$\boldsymbol{\mu} = \mathbf{K}_{NM}\mathbf{K}_M^{-1}\mathbf{m} \tag{9}$$

$$\boldsymbol{\Sigma} = \mathbf{K}_N - \mathbf{K}_{NM}\mathbf{K}_M^{-1}(\mathbf{K}_M - \mathbf{S})\mathbf{K}_M^{-1}\mathbf{K}_{MN} \tag{10}$$

このようにして得られた変分下限は

$$\mathcal{L} = \sum_{i=1}^{N} \left\{\mathbb{E}_{q(f(\mathbf{x}_i))}\left[\log p(y_i|f(\mathbf{x}_i))\right] - \frac{1}{N}\mathrm{KL}(q(\mathbf{u})\|p(\mathbf{u}))\right\} \tag{11}$$

となり，サンプル点ごとの和で表されることがわかる．したがって変分下限の最大化は，ニューラルネットワークにおけるコストの最小化と同様に，確率的勾配降下法に基づくパラメータの学習が可能である．

具体的に式 (11) の右辺を見ると，第一項は予測分布の出力変数への適合度を示している．一方で，第二項は補助変数の変分分布と事前分布との KL ダイバージェンスであり，正則化項と見なすことができる．SVGP の学習時には，補助点 $\mathbf{Z}$ および

(注1)：SVGP という略称はガウス過程のツールキットである GPy [10] に基づいている．

その出力変数の変分分布 $q(\mathbf{u})$，さらにカーネル関数のパラメータをミニバッチ学習により最適化する．予測時には式 (8) と同様の変分分布を用いて次の式より予測分布 $p(\mathbf{y}_T)$ を推定する．

$$p(\mathbf{y}_T) = \int p(\mathbf{y}_T|\mathbf{f}_T)q(\mathbf{f}_T)d\mathbf{f}_T \tag{12}$$

$$q(\mathbf{f}_T) = \mathcal{N}(\mathbf{f}_T; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{13}$$

$$\boldsymbol{\mu} = \mathbf{K}_{TM}\mathbf{K}_M^{-1}\mathbf{m} \tag{14}$$

$$\boldsymbol{\Sigma} = \mathbf{K}_T - \mathbf{K}_{TM}\mathbf{K}_M^{-1}(\mathbf{K}_M - \mathbf{S})\mathbf{K}_M^{-1}\mathbf{K}_{MT} \tag{15}$$

3. SVGP に基づく音声合成

3.1 コンテキストのためのカーネル関数

本節では SVGP をパラメトリック音声合成に用いることを考える．ガウス過程回帰においてカーネルの設計は非常に重要な問題であるが，本研究では文献 [4] で提案したカーネルを修正したものを用いる．カーネルの定義を以下に示す．

$$\kappa(\mathbf{x}_m, \mathbf{x}_n) = \sum_{k=1}^{K} \theta_{r,k}^2 \bar{\kappa}_k(\mathbf{x}_{m,k}, \mathbf{x}_{n,k}) + \delta_{mn}\theta_{\text{floor}}^2 \tag{16}$$

$$\bar{\kappa}_k(\mathbf{x}_{m,k}, \mathbf{x}_{n,k}) = \frac{\kappa_k(\mathbf{x}_{m,k}, \mathbf{x}_{n,k})}{\sqrt{\kappa_k(\mathbf{x}_{m,k}, \mathbf{x}_{m,k})\kappa_k(\mathbf{x}_{n,k}, \mathbf{x}_{n,k})}} \tag{17}$$

$$\kappa_k(\mathbf{x}_{m,k}, \mathbf{x}_{n,k}) = \sum_{u=-1}^{+1}\sum_{v=-1}^{+1}\Big[w\left(\mathbf{p}_{m,k}^{(u)}\right) w\left(\mathbf{p}_{n,k}^{(v)}\right) \cdot \kappa_{kp}\left(\mathbf{p}_{m,k}^{(u)}, \mathbf{p}_{n,k}^{(v)}\right)\kappa_{kc}\left(\mathbf{c}_{m,k}^{(u)}, \mathbf{c}_{n,k}^{(v)}\right)\Big] \tag{18}$$

$k = (1, \ldots, K)$ 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のインデックスを表し，その部分カーネル $\bar{\kappa}_k(\mathbf{x}_{m,k}, \mathbf{x}_{n,k})$ を重み $\theta_{r,k}^2$ で加算したものをカーネルとする．それぞれの部分カーネルの対角成分を 1 にするため，式 (17) の処理によりカーネルの正規化 [14] を行っている．部分カーネル $\kappa_k(\mathbf{x}_{m,k}, \mathbf{x}_{n,k})$ のコアの部分は，イベントからの相対位置の類似度を表すカーネル $\kappa_{kp}(\cdot)$ と音素弁別特性やモーラの高低など音声単位のコンテキストの類似度を表すカーネル $\kappa_{kc}(\cdot)$ である．例えば音素のイベントに対する部分カーネルは，音素が類似していて，さらに音素内位置が似ている場合にカーネル関数の値が大きくなるように定義している．添字 $(-1), (0), (+1)$ および重み関数 $w(\cdot)$ は，カーネル関数が音素などの境界で不連続にならないために用いられている [3].

3.2 音響特徴量の尤度関数

提案手法では音響特徴量ごとに個別の尤度関数を用いる．メルケプストラムや非周期性指標などの連続値に対しては，一般的なガウス過程回帰と同様に以下の正規分布を尤度関数に用いる．

$$p(y|f(\mathbf{x})) = \mathcal{N}(y; f(\mathbf{x}), \sigma_\nu^2) \tag{19}$$

ただし，σ_ν^2 はノイズの分散を表すハイパーパラメータである．F0 系列は無声区間を含むため，対数 F0 の尤度関数は HMM 音声合成の枠組みで用いられる多空間確率分布 (MSD) を使用する．

表 1　提案法における時系列イベント．AP および BG はそれぞれアクセント句，呼気段落を示す．
Table 1　Temporal events used in the proposed method. The units AP and BG denote accent phrase and breath group, respectively.

k	単位	位置	イベント特徴量
1	音素	開始	音素弁別特性,
2	音素	終了	継続長
3	モーラ	開始	アクセントの高／低，継続長,
4	モーラ	終了	モーラ位置
5	AP	開始	アクセント型,
6	AP	アクセント核の終了	句の位置,
7	AP	終了	モーラ数，継続長
8	BG	開始	句の位置,
9	BG	終了	モーラ数，継続長
10	発話	開始	文末語，疑問調
11	発話	終了	モーラ数，継続長

$$p(y|f(\mathbf{x})) = \begin{cases} \mathcal{N}(y; f(\mathbf{x}), \sigma_\nu^2), & \text{if voiced} \\ 1, & \text{if unvoiced} \end{cases} \tag{20}$$

なお，MSD を尤度関数として使用することは，GPR において有声区間のみでのモデル化を行うことと等価である．有声／無声フラグのモデル化には，ガウス過程分類の枠組みで用いられる標準正規分布の累積分布関数

$$p(y|f(\mathbf{x})) = \Phi(yf(\mathbf{x})) \tag{21}$$

を用いる．ただし，有声のときは $y = 1$，無声のときは $y = -1$ とする．

提案手法では，複数の音響特徴量を共通のガウス過程を用いて同時にモデル化する．具体的には，カーネル関数のパラメータおよび補助点 $\mathbf{Z}$ は音響特徴量間で共有し，変分分布 $q(\mathbf{u})$ は音響特徴量ごとに個別に学習する．

4. GP-DNN ハイブリッドモデルに基づく音声合成

これまでの GPR 音声合成 [3,4] では，13 次元の音素弁別特性などに代表されるように，低次元の特徴量を入力変数として使用していた．しかし，例えばトライフォン同士の類似度を定義するカーネル関数において，音素弁別特性を入力変数に用いることは必ずしも最適とは限らず，適切な入力特徴量の決定が重要な問題である．一方で，DNN 音声合成の枠組みでは，コンテキストに対する質問の回答を示したバイナリ値を含む数百次元のベクトルを入力変数として使用している．DNN は深層構造によって自動的に特徴表現を得られる手法として知られている [15] が，DNN 音声合成において入力に高次元のベクトルを用いる理由の一つとして，コンテキストの特徴表現を効率的

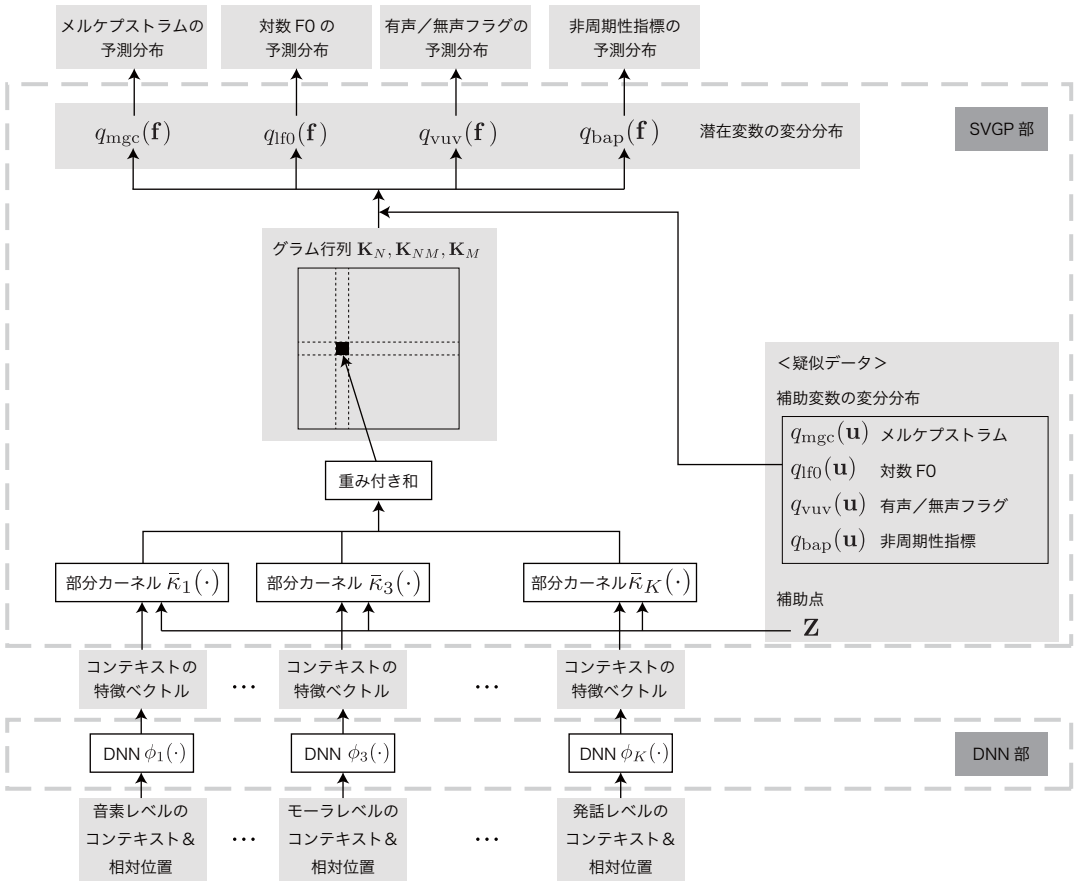

図 1　GP-DNN ハイブリッドモデルに基づく音響モデリングの枠組み
Fig. 1　An outline of acoustic feature modeling based on GP-DNN hybrid model.

に獲得できるという想定が挙げられる．ここで，この高次元のベクトルをガウス過程の枠組みに組み込むことを考える．しかし，ガウス過程において数百次元ものベクトルを入力変数として使用した場合，次元の呪いによりしばしばグラム行列がスパースになり，学習が困難になる可能性がある．

この問題に対し，本研究では GP-DNN ハイブリッドモデル [8] を SVGP 音声合成に導入する．提案法の概要図を図 1 に示す．GP-DNN ハイブリッドモデルは SVGP の入力変数に DNN の出力を用いる方法であり，ガウス過程が不得手とする高次元ベクトルからの特徴抽出を DNN で補うことを目的としている．提案法では，式 (18) の部分カーネルにおいて位置カーネル $\kappa_{kp}(\cdot)$ およびイベント特徴量カーネル $\kappa_{kc}(\cdot)$ を組み合わせた $\kappa_{k\phi}(\cdot)$ で置き換え，以下のカーネル関数とする．

$$\kappa'_k(\mathbf{x}_{m,k}, \mathbf{x}_{n,k}) = \sum_{u=-1}^{+1} \sum_{v=-1}^{+1} \Big[w\left(\mathbf{p}_{m,k}^{(u)}\right) w\left(\mathbf{p}_{n,k}^{(v)}\right) \cdot \kappa_{k\phi}\left(\boldsymbol{\phi}_k(\mathbf{p}_{m,k}^{(u)}, \mathbf{c}_{m,k}^{(u)}), \boldsymbol{\phi}_k(\mathbf{p}_{n,k}^{(v)}, \mathbf{c}_{n,k}^{(v)})\right) \Big] \tag{22}$$

ただし，$\boldsymbol{\phi}_k(\cdot)$ はフィードフォワード型のニューラルネットワークであり，コンテキストの特徴ベクトルの抽出を行うことを想定している．GP-DNN ハイブリッドモデルは上位層に SVGP を，下位層に DNN を用いており，この階層構造のネットワークは最下層まで微分可能である．したがって変分下限からのバックプロパゲーションによって DNN のパラメータと SVGP のパラメータは同時学習が可能である．

GP-DNN を DNN と比較すると，DNN が値を推定するのに対し GP-DNN は予測分布を推定する．例えば未知のコンテキストに対しては大きい予測分散，すなわち “不確かさが高い” という予測結果を出力する．したがって，未知のコンテキストに対して予測範囲外の値を出力する可能性のある DNN に比べ，頑健性の高い予測が期待できる．また，変分下限をコスト関数としているため DNN と比較すると過学習の起こりにくいモデルといえる．

5. 実　　験

5.1 実験条件

データベースには音声合成システム XIMERA [16] 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．用いられた文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 を音響特徴量として使用した．

合成時には，動的特徴量からのパラメータ生成 [17] を行い．主観評価には GV [18] を制約として生成したメルケプストラムを用いた．すべての入力変数および連続値の出力変数に対して

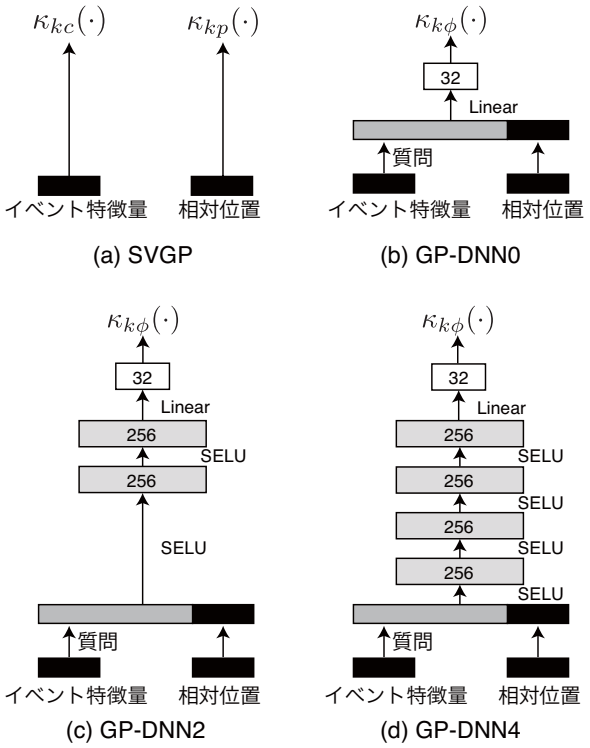

図 2　実験における SVGP および GP-DNN の特徴抽出の比較

Fig. 2　A comparison among extraction methods of a feature vector that is used as an input of the kernel function of k-th temporal event.

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．また学習時のパラメータ最適化には ADAM [19] を使用し，学習係数は 0.001 とした．

GP-DNN ハイブリッドモデルに用いた時系列イベントを表 1 に示す．単位ごとにイベント特徴量を抽出し，音素，モーラ，アクセント句，呼気段落および発話のイベント特徴量の次元はそれぞれ，243，82，136，76，35 となった．相対位置情報には 6 次元の特徴量を用いた．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．提案法では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 [20] を活性化関数として使用した．カーネル関数 $\kappa_{kp}(\cdot)$,$\kappa_{kc}(\cdot)$,$\kappa_{k\phi}(\cdot)$ には RBF カーネルを使用し，補助点数 M は 1024 とした．

実験では次に示す 5 手法の比較を行った．GP-DNN および SVGP における入力特徴量の違いを図 2 に示す．

- GP-DNN

提案法である SVGP と DNN のハイブリッド手法．GP-DNN2 と GP-DNN4 はそれぞれ隠れ層の数が 2 および 4 であることを示す．GP-DNN0 は隠れ層を用いず，線形変換のみでカーネル関数の入力特徴量を抽出する手法である．最適化におけるミニバッチサイズは 1024 とした．

- SVGP

3 節で説明した SVGP に基づく手法．GP-DNN とは異なり，コンテキスト特徴量を質問を適用せず直接カーネル関数の入力とする．

- PIC-GP

表 2　合成音声の原音声に対する音響特徴量歪．MCEP: メルケプストラム距離 [dB]，F0: 対数 F0 の RMSE[cent]，V/UV: 有声／無声誤り率 [%]，BAP: 非周期性指標歪 [dB]，DUR: 音素継続長歪 [ms]．

Table 2　The acoustic feature distortions between original and synthetic speech. MCEP: mel-cepstral distortion [dB], F0: RMSE of log F0 [cent], V/UV: V/UV error rate [%], BAP: aperiodicity distortion [dB], DUR: RMSE of phone duration [ms].

手法	MCEP	F0	V/UV	BAP	DUR
PIC-GP	5.11	181	5.11	3.28	16.4
DNN3	5.07	179	4.83	3.24	16.9
DNN6	5.02	179	4.82	3.23	18.1
LSTM	4.92	178	4.75	3.21	17.4
SVGP	5.04	178	4.78	3.28	15.7
GP-DNN0	5.24	210	5.06	3.38	16.2
GP-DNN2	5.01	184	5.08	3.27	16.4
GP-DNN4	4.93	180	4.62	3.24	16.3

ガウス過程の近似手法として SVGP ではなく従来の PIC 近似を用いる [4]．PIC 近似におけるブロック分割には HMM 音声合成で用いられる決定木を使用し，各ブロックにおける最大フレーム数 1024 とした．また，ハイパーパラメータ最適化には EM アルゴリズムに基づく手法 [17] を用いた．

- DNN

文献 [1] に基づく DNN 音声合成．主なネットワークの設定は文献 [2] に基づき，活性化関数にはハイパボリックタンジェント関数，隠れ層のノード数は 1024，隠れ層の数は 3 および 6 とし，それぞれの手を DNN3，DNN6 とした．

- LSTM

LSTM-RNN に基づく手法 [2]．文献 [2] と同様に 2 層のフィードフォワード層と 2 層の双方向 LSTM 層から成るネットワークを用いた．ミニバッチサイズは 16 とし，隠れ層のノード数は 1024 とした．

5. 2　客観評価結果

客観評価として，合成音声と原音声の音響特徴量歪を評価した．表 2 に結果を示す．まず，GP-DNN ハイブリッドモデルの GP-DNN0, GP-DNN2, GP-DNN4 を比較すると，隠れ層が多くなるとメルケプストラム距離が小さくなるという結果を得た．また，GP-DNN4 では DNN を使用しない SVGP や従来法の PIC-GP に比べメルケプストラム距離が小さくなった．このことは，DNN が特徴抽出器として効果的に働いたことを示している．さらに，GP-DNN4 はネットワークにリカレント構造を持っていないが，LSTM-RNN と同程度のスペクトル歪となっている．一方で．F0 や音素継続長の歪は GP-DNN4 より SVGP の方が小さく，この結果から，深層構造は F0 や継続長のモデル化には必ずしも必要でないことがわかる．

表 3　対比較による主観評価結果 [%].
Table 3　Results of paired comparison test [%].

GP-DNN4	SVGP	PIC	Neutral	p	Z
58.3	6.7		35.0	$< 10^{-4}$	9.14
75.8		3.3	20.8	$< 10^{-4}$	15.40
	51.7	13.3	35.0	$< 10^{-4}$	5.92
GP-DNN4	DNN6	LSTM	Neutral	p	Z
63.3	13.3		23.3	$< 10^{-4}$	7.62
32.5		25.8	41.7	0.34	0.96

5.3　主観評価結果

主観評価実験では対比較による合成音声の評価を行った．被験者は 6 人で，各被験者にはテストデータからランダムに選ばれた 10 文章に対し，合成音声のペアを聴きそのどちらが自然かを選択するよう指示を与えた．また両者に差がないと感じられた場合には “neutral” を選べるものとした．

結果と p 値および Z 値を表 3 に示す．表から，SVGP と PIC-GP を比較した場合 SVGP の方が有意に高いスコアを得た．この理由としては，PIC-GP では HMM に基づく決定木を用いており，決定木による予測性能の影響を受けてしまうことが考えられる．また GP-DNN4 と SVGP を比較すると，GP-DNN4 のスコアが高く，スペクトル歪の客観評価と同様に DNN による特徴抽出が効果的であることがわかる．さらに，GP-DNN4 と他の深層構造に基づく手法である DNN および LSTM-RNN と比べると，GP-DNN4 は DNN6 に比べ有意にスコアが高く，LSTM-RNN と同程度であるという結果を得た．

6. むすび

本稿では，確率的変分ガウス過程 (SVGP) の入力に DNN の出力を用いる GP-DNN ハイブリッドモデルに基づく音声合成手法の提案の行った．GP-DNN モデルでは，変分下限がサンプル点の和で表され，DNN のパラメータまで微分可能なネットワークで表されるため，一般的な DNN の枠組みと同様に，確率的勾配降下法に基づく最適化が可能である．主観評価実験から，提案法はフィードフォワード型の DNN や従来の PIC 近似を用いたガウス過程に比べ自然性が有意に高く，また，提案法はリカレント構造を持っていないにも関わらず LSTM-RNN と同程度の自然性という結果を得た．今後の課題として，多様なデータベースを用いた実験や，提案法にリカレント構造を導入することによる品質の向上が挙げられる．

7. 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] Y. Fan, Y. Qian, F. Xie, and F.K. Soong, “TTS synthesis with bidirectional LSTM based recurrent neural networks,” Proc. INTERSPEECH, pp.1964–1968, 2014.
[3] 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.
[4] T. Koriyama and T. Kobayashi, “Prosody generation using frame-based Gaussian process regression and classification for statistical parametric speech synthesis,” Proc. ICASSP, pp.4929–4933, 2015.
[5] J. Hensman, N. Fusi, and N. Lawrence, “Gaussian processes for big data,” Proc. AUAI, pp.282–290, 2013.
[6] J. Hensman, A.G.d.G. Matthews, and Z. Ghahramani, “Scalable variational Gaussian process classification,” Proc. AISTATS, pp.1648–1656, 2015.
[7] A.G. Wilson, Z. Hu, R. Salakhutdinov, and E.P. Xing, “Stochastic variational deep kernel learning,” Proc. NIPS, pp.2586–2594, 2016.
[8] J. Bradshaw, A.G.d.G. Matthews, and Z. Ghahramani, “Adversarial examples, uncertainty, and transfer testing robustness in Gaussian process hybrid deep networks,” arXiv preprint arXiv:1707.02476, 2017.
[9] H. Salimbeni and M. Deisenroth, “Doubly stochastic variational inference for deep gaussian processes,” Proc. NIPS, pp.4591–4602, 2017.
[10] GPy, “GPy: A gaussian process framework in python,” http://github.com/SheffieldML/GPy.
[11] M. Seeger, C. Williams, and N. Lawrence, “Fast forward selection to speed up sparse Gaussian process regression,” Proc. AISTATS, 2003.
[12] E. Snelson and Z. Ghahramani, “Sparse Gaussian processes using pseudo-inputs,” Proc. NIPS, pp.1257–1264, 2006.
[13] E. Snelson and Z. Ghahramani, “Local and global sparse Gaussian process approximations,” Proc. AISTATS, pp.524–531, 2007.
[14] C.E. Rasmussen and C.K.I. Williams, Gaussian Processes for Machine Learning, The MIT press, 2006.
[15] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016. http://www.deeplearningbook.org.
[16] 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.2688–2698，2006.
[17] 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.
[18] T. Toda and K. Tokuda, “A speech parameter generation algorithm considering global variance for HMM-based speech synthesis,” IEICE Trans. Inf. & Syst., vol.E90-D, no.5, pp.816–824, 2007.
[19] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.
[20] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, “Self-normalizing neural networks,” arXiv preprint arXiv:1706.02515, 2017.