論文 / 著書情報
Article / Book Information

論題(和文)	GPR音声合成における深層構造の利用の検討
Title(English)	
著者(和文)	郡山知樹, 小林隆夫
Authors(English)	Tomoki Koriyama, Takao Kobayashi
出典(和文)	日本音響学会2018年春季研究発表会講演論文集, Vol. , No. , pp. 1507-1508
Citation(English)	, Vol. , No. , pp. 1507-1508
発行日 / Pub. date	2018, 3

GPR 音声合成における深層構造の利用の検討*

○郡山知樹, 小林隆夫 (東工大)

1 はじめに

ガウス過程回帰 (GPR) に基づく音声合成は，カーネル回帰のベイズ推定の枠組みで，コンテキスト同士の相関関係を用いて音声パラメータを生成する手法である [1]．GPR 音声合成においてカーネルの設計は重要な課題であるが，例えばトライフォンのようなカテゴリカルなコンテキストの類似度を表現するカーネルを，人手で設計することは困難である．近年，ガウス過程において深層構造を用いて入力変数を変換することによって，カーネルの構築を行う手法が提案されている [2, 3]．これまでの報告 [4] で我々は，DNN を用いてコンテキストの特徴表現を抽出し，それを GPR の入力とする GP-DNN ハイブリッドモデル [2] の有効性を示した．本研究では，GP-DNN ハイブリッドモデルだけでなく，ガウス過程自体を深層構造に適用した深層ガウス過程 [3] を用いた音響モデリング手法を提案し，GPR 音声合成における深層構造の有用性を評価する．

2 GP-DNN ハイブリッドモデル

GPR 音声合成では，データ点 $i(i = 1 \dots N)$，次元 d の音響特徴量 y_i^d のノイズ成分を除いた潜在関数変数 $f_i^d = f^d(\mathbf{x}_i)$ に対し，潜在関数 $f^d(\cdot)$ がガウス過程 $\mathcal{GP}(m^d(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ に従うと仮定する．ただし，$\mathbf{x}$ はフレームレベルのコンテキスト，$m^d(\mathbf{x})$ は平均関数，$k(\mathbf{x}, \mathbf{x}')$ はコンテキスト同士の関係を表すカーネル関数である．

GP-DNN ハイブリッドモデルに基づく音声合成 [4] では，Fig. 1(b) に示すように，コンテキストをフィードフォワード型の DNN の入力とし，得られた低次元特徴量 $\boldsymbol{\phi}(\mathbf{x})$ をカーネル関数の入力とする．このとき，確率的変分ガウス過程 (SVGP)[5] をガウス過程の近似として用いると，以下の対数周辺尤度の変分下限 $\mathcal{L}$ を最大化することで GP-DNN ハイブリッドモデルの学習が行われる．

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{d=1}^{D} \left\{ \mathbb{E}_{q(f_i^d)} \left[\log p(y_i^d | f_i^d) \right] - \frac{1}{N} \mathrm{KL}(q(\mathbf{u}^d) \| p(\mathbf{u}^d)) \right\} \quad (1)$$

$$q(f_i^d) = \mathcal{N}(f_i^d; \mu_i^d, \Sigma_i^d) \quad (2)$$

$$\mu_i^d = \mathbf{K}_{iM} \mathbf{K}_M^{-1} \mathbf{m}^d \quad (3)$$

$$\Sigma_i^d = \mathbf{K}_i - \mathbf{K}_{iM} \mathbf{K}_M^{-1} (\mathbf{K}_M - \mathbf{S}^d) \mathbf{K}_M^{-1} \mathbf{K}_{Mi} \quad (4)$$

ただし，$\mathbf{K}_{iM}, \mathbf{K}_i, \mathbf{K}_M$ は $k(\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}'))$ を要素に持つグラム行列，$q(\mathbf{u}^d)$ は補助変数 $\mathbf{u}^d$ の変分分布でありガウス分布 $\mathcal{N}(\mathbf{u}^d; \mathbf{m}^d, \mathbf{S}^d)$ で表される．GP-DNN ハイブリッドモデルでは，誤差逆伝播法により SVGP のモデルパラメータと DNN のパラメータを同時に学習できる．また，変分下限がデータ点ごとの和となっているため，確率的勾配降下法に基づく最適化が可能である．音声合成時には，式 (2) を用いて音声パラメータの予測分布を推定し，これを用いて音声パラメータ生成を行う．

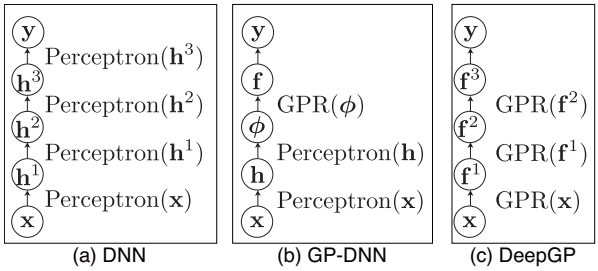

Fig. 1 DNN および提案法のモデル構造

3 深層ガウス過程

GP-DNN ハイブリッドモデルでは DNN の深層構造によってコンテキストの特徴表現を抽出可能であることを想定しているが，重みの事前分布を考慮していないため DNN と同様に重みのパラメータが過学習しやすい．これに対し，Fig. 1(c) に示すような，GPR を階層構造に接続した深層ガウス過程 (GP)[6] が提案されている．深層 GP は，層 l，次元 d の隠れ関数 $f_i^{l,d}$ $(d = 1 \dots D_l)$ が D_{l-1} 次元の層 $l-1$ の隠れ関数変数 $\mathbf{f}_i^{l-1}$ を入力とするガウス過程に従って生成されることを仮定したモデルである．深層 GP では，すべての層で GPR によるベイズ推定を行うため，GP-DNN に比べ過学習しにくいモデルといえる．

しかし深層 GP を計算量の問題から音声合成に直接適用することが困難である．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) に基づく深層 GP[3] を用いる．DSVI では SVGP と同様に対数周辺尤度の変分下限を以下のようなサンプル点ごとの値の和の形式で表す．

$$\mathcal{L} = \sum_{i=1}^{N} \left\{ \sum_{d=1}^{D_L} \mathbb{E}_{q(f_i^{L,d})} \left[\log p(y_i^{L,d} | f_i^{L,d}) \right] - \frac{1}{N} \sum_{l=1}^{L} \sum_{d=1}^{D_l} \mathrm{KL}(q(\mathbf{u}^{l,d}) \| p(\mathbf{u}^{l,d})) \right\} \quad (5)$$

ただし $q(\mathbf{u}^{l,d})$ は層 l，次元 d の補助変数 $\mathbf{u}^{l,d}$ の変分分布であり，ガウス分布 $\mathcal{N}(\mathbf{u}^{l,d}; \mathbf{m}^{l,d}, \mathbf{S}^{l,d})$ で表される．

変分事後分布 $q(\mathbf{f}_i^L)$ は

$$q(\mathbf{f}_i^L) = \int \prod_{l=1}^{L} q(\mathbf{f}_i^l | \mathbf{f}_i^{l-1}) d\mathbf{f}_i^l \quad (6)$$

となるが，この積分は一部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 では，下位層のサンプル点 $\hat{\mathbf{f}}_i^{l-1}$ から上位層の隠れ変数 $\hat{\mathbf{f}}_i^l$ を条件付き分布 $q(\mathbf{f}_i^l | \hat{\mathbf{f}}_i^{l-1})$ を用いて再帰的にサンプリングする．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．

$$q(\mathbf{f}_i^L) = \frac{1}{S} \sum_{s=1}^{S} q(\mathbf{f}_i^L | \hat{\mathbf{f}}_{i,s}^{L-1}) \quad (7)$$

---

*On the Use of Deep Model Architecture for GPR-based Speech Synthesis. by KORIYAMA, Tomoki, KOBAYASHI, Takao (Tokyo Institute of Technology)

Table 1　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]，f_o: 対数 f_o 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])

手法	MCEP	f_o	V/UV	BAP	DUR
DNN(ReLU)	4.80	175	4.70	3.10	16.4
DNN(tanh)	5.21	193	5.42	3.26	19.1
LSTM	**4.71**	183	4.65	**3.08**	18.0
GP-DNN	4.78	175	4.55	3.13	17.4
DeepGP	4.74	**174**	**4.49**	3.09	**15.1**

これを式 (5) に代入すると，変分下限はデータ点とサンプリングによって得られた点の二重の和の形式で表される．したがって，SVGP と同様に DSVI に基づく深層 GP では，確率的勾配降下法に基づくパラメータ最適化が可能である．合成時においても学習時と同様に，予測変分分布 $q(\mathbf{f}_i^L)$ を解析的に求められない．本研究では，各層の予測平均 $\mathbb{E}[q(\mathbf{f}_i^l|\hat{\mathbf{f}}_i^{l-1})]$ を $\hat{\mathbf{f}}_i^l$ として予測分布 $q(\mathbf{f}_i^L)$ を近似する．

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[7] 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 毎に STRAIGHT を用いて f_o，スペクトル包絡，非周期性指標を抽出し，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 と有声／無声フラグを音響特徴量として使用した．継続長モデルでは音素継続長を音響特徴量とした．

音声単位ごとにコンテキストを抽出し，音素，モーラ，アクセント句，呼気段落および発話のコンテキストの次元を 249，88，142，82，41 とした．GP-DNN ハイブリッドモデルおよび，深層 GP の最上位層のカーネルには文献 [4] と同様に音声イベントごとのカーネルの和を求める加算構造のカーネルを用いる．SVGP の補助点数は 1024 とした．またパラメータ最適化手法には Adam を用いた．

GP-DNN ハイブリッドモデル (**GP-DNN**) の DNN の隠れ層は 4 層，256 素子とし，出力層は 32 次元とした．活性化関数には，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[8] を使用した．深層 GP(**DeepGP**) では層の数を 6 とし，最上位層以外の層では補助点数は 256，次元は 32，カーネルは RBF カーネルとした．

比較手法として DNN 音声合成 (**DNN**)[9] と双方向 LSTM に基づく RNN 音声合成 (**LSTM**)[10] を用いた．DNN 音声合成では隠れ層は 6 層，ノード数は 1024 とし，活性化関数には ReLU および tanh を用いた．RNN 音声合成では下位 2 層のフィードフォワード層と上位 2 層の双方向 LSTM 層からなるネットワークを使用した．活性化関数は tanh，隠れ層の素子数は 512 とした．

4.2　結果

客観評価結果として合成音声の原音声に対する歪を Table 1 に示す．メルケプストラムや非周期性指標は LSTM において最も歪が小さくなったが，対数 f_o，

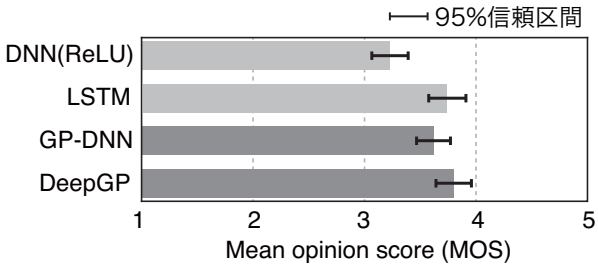

Fig. 2　主観評価結果

有声／無声フラグ，音素継続長の歪は深層 GP が最も小さいという結果を得た．また表から，DNN 音声合成や GP-DNN ハイブリッドモデルに比べ深層 GP はすべての特徴量で歪が小さいことがわかる．

主観評価実験では MOS 試験による自然性の比較評価を行った．被験者は 7 名で各被験者は評価データ 60 文の中からランダムに選ばれた 15 文を評価した．合成音声の自然性を 5 段階で評価し，その MOS 値を求めた．結果を Fig. 2 に示す．DNN 音声合成と他手法を比較すると，$\alpha = 0.05$ で有意に DNN の MOS の値が低かった．また，GP-DNN ハイブリッドモデルと深層 GP は，モデルにリカレント構造を有していないにも関わらず，LSTM-RNN に基づく音声合成と同程度のスコアを得た．

5　おわりに

本稿では GPR 音声合成に深層構造を導入した手法として，GP-DNN ハイブリッドモデルと深層ガウス過程に基づく手法を提案した．提案法は DNN 音声合成より有意に自然性が高く，LSTM-RNN 音声合成と同程度の自然性という主観評価結果となった．提案法へのリカレント構造の導入や様々なデータでの評価が今後の課題である．

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

参考文献

[1] 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.
[2] Bradshaw et al., “Adversarial examples, uncertainty, and transfer testing robustness in Gaussian process hybrid deep networks,” arXiv:1707.02476, 2017.
[3] Salimbeni et al., “Doubly stochastic variational inference for deep Gaussian processes,” Proc. NIPS, pp.4591–4602, 2017.
[4] 郡山 他，“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] Hensman et al., “Scalable variational Gaussian process classification,” Proc. AISTATS, pp.1648–1656, 2015.
[6] Damianou et al., “Deep Gaussian processes,” Proc. AISTATS, pp.207–215, 2013.
[7] 河井 他，“大規模コーパスを用いた音声合成システム XIMERA，” 信学論 (D)，89(12)，pp. 2688–2968, 2006.
[8] Klambauer et al., “Self-normalizing neural networks,” Proc. NIPS, pp.972–981, 2017.
[9] Zen et al., “Statistical parametric speech synthesis using deep neural networks,” Proc. ICASSP, pp.7962–7966, 2013.
[10] Fan et al.,, “TTS synthesis with bidirectional LSTM based recurrent neural networks,” Proc. INTERSPEECH, pp.1964–1968, 2014.