様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月　 ６ 日現在

機関番号：１２６０１
研究種目：若手研究
研究期間：2019～2021
課題番号：１９Ｋ２０２９２

研究課題名（和文）深層ガウス過程に基づく統計的音声合成

研究課題名（英文）A Study of Deep Gaussian Process Based Statistcal Speech Synthesis

研究代表者
郡山　知樹（Koriyama, Tomoki）

東京大学・大学院情報理工学系研究科・講師

研究者番号：５０７４９１２４

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：未知のデータに頑健なモデルとして提案されている深層ガウス過程(Deep Gaussian process, DGP)を，時系列のモデル化が必要な音声合成に応用する手法として，リカレント構造やself-attention構造，sequence-to-sequence構造を持つDGPを提案した．提案手法は同様の構造を持つDNN音声合成より高品質な音声を合成できる傾向があることを示した．本研究課題の成果によって，ニューラルネットワークに用いられる様々な構造がDGPでも実現可能であり，ベイズの特徴を用いた頑健な深層学習ができることを示した．

研究成果の学術的意義や社会的意義
現在，多くの機械学習の研究はDNNを基盤要素としているが，DNNの学習におけるハイパーパラメータの調整は手間のかかるものであり，機械学習モデルの構築は職人作業のようになっている現状がある．代替となり得るモデルとしてガウス過程回帰に注目が集まっているが，自由度が低く様々なデータに適用できない問題があった．本研究の応用実験によってガウス過程回帰の深層モデルとしての自由度の向上を明らかにした．この成果によって，音声に限らず自由度の高い深層学習モデルの頑健な学習への道筋の一つを示した．

研究成果の概要（英文）：We proposed an extension of deep-Gaussian-process (DGP)-based speech synthesis to enable time-series modeling of speech characteristics. Specifically, we proposed DGP with recurrent, self-attention, and sequence-to-sequence architecture. The proposed speech synthesis methods tend to generate more natural-sound speech than that generated by DNN-based ones that have similar architectures of DGP. The results of this research project show that various structures used in DGP can be used in a similar way to DNN, and that robust deep learning using Bayesian features is possible.

研究分野：音声情報処理

キーワード：ガウス過程　深層学習　音声合成　潜在変数モデル　時系列モデル

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

人工知能ブームの基盤技術である，深層ニューラルネットワーク(以下 DNN とする)は，変数の線形変換と非線形変換から成るパーセプトロンを多重に組み合わせた深層構造によって，任意の関数を近似する機械学習モデルであり，画像処理や言語処理，音声情報処理を始めとして様々な分野で DNN が活用されている．スマートスピーカーや，スマートフォンにおける音声対話システムにおいて重要な要素である音声合成に対しても，DNN は基盤技術の一つであり，広く研究されている.このとき音声合成では，入力であるテキストから，出力である音声波形へのマッピングを行う関数を学習する．具体的には音声の短時間の音韻情報や韻律情報を入力として，音声波形の生成に用いられる音響特徴量を出力する音声フレームレベルの DNN を学習するのが伝統的なシステムである．

DNN は大量のパラメータに由来する高い表現力を持ち，さらに任意の微分可能な関数を組み合わせることで表現力をさらに高めることが出来る．音声合成においては，時系列に沿った連続的な変化を表現するリカレント構造，短期間の特徴を獲得する畳み込み構造，時系列間の類似度を表現する注意(attention)機構などが用いられる．一方で，DNN の学習で得られた関数は，未知のデータに対して頑健であることを保証しないため，未知の入力に対して全く予想できない結果を返すことがある．学習データを大量に用意することで，一般的にこの問題は解消されるが，音声合成において大量にデータを用意することは必ずしも容易ではない．

そこで，本研究では DNN に代わるモデル深層ガウス過程(deep Gaussian process, DGP)を用いた音声合成に着目する．DGP はカーネル法に基づく手法であるため，未知のデータに対しても学習データと類似したデータから音声を生成する．また，DGP はベイズモデルであるため，モデルの複雑さを考慮した学習が可能になる．これらによって音声の頑健性が期待でき，実際にフィードフォワード型の単純な深層構造では DGP 音声合成が DNN 音声合成より自然な音声を合成できることをこれまでに報告している．

２．研究の目的

本研究の主な目的は DGP 音声合成の時系列表現獲得のためのモデル拡張である．これまでの DGP 音声合成は単純なフィードフォワード型だけであったが，DNN 音声合成で用いられるようなリカレント型や注意機構を機能として組み込むことの有効性を評価する．同時に，音声に限らず DGP 自体の DNN と同様の拡張性が示されていないため，DGP のモデルとしての拡張性を評価する．

また，DGP 音声合成の応用先として，ベイズモデルの特長を活用した潜在変数の自動獲得がある．この応用から話者やアクセントなどの特徴を自動獲得し，より自由度の高い音声合成システムへの拡張可能性を検討する．

3. 研究の方法

(1) 時系列を考慮可能な構造への深層ガウス過程の拡張

音声合成はテキストの文字列から音声パラメータの時系列を予測するシステムであるため，音声フレームレベルで予測を行うフィードフォワード型は性能に限界がある. 実際に DNN 音声合成では，フィードフォワード型でなくリカレント構造や注意機構構造のように系列を考慮したモデルが研究の主流になっている. 本研究では，時系列情報を深層ガウス過程でモデル化できるように，図 2 に示すリカレント型，畳み込み型，注意機構型といった，DNN の枠組みで有効性の確認されている構造を深層ガウス過程で実現する方法を検討する. 具体的には，DNN でパーセプトロンになっている部分をガウス過程に置き換える方針で，モデルの設計を行う. このとき，ガウス過程の計算量はパーセプトロンより多いため，単純に置き換えるだけでは大量の計算時間を必要としてしまう. そこで並列計算の効率を考慮したモデルの設計を工夫する.

(2) 潜在変数モデルを用いた話者・スタイルの獲得

学習データの音声が十分に得られなくても，目標とする話者や感情表現などの発話スタイルのモデルを構築する手法として，既存の音声データから学習した平均声モデルから目標話者・スタイルのモデルを学習するモデル適応がある.本研究では，モデル適応を深層ガウス過程の枠組みで実現することを考える.具体的には，話者や感情表現を低次元の潜在変数で表現し，少量の適応データを潜在変数の推定に使用する方法を用いる.深層ガウス過程では確率モデルの特性を用いることで自然に潜在変数モデルの導入が可能である.同様の手法は DNN に基づく潜在変数モデルでも実現可能であるが，深層ガウス過程では潜在変数同士の類似度を用いるカーネル法に基づく手法であることから，提案法では 2 話者の音声の補間や感情の表出度合いなど，類似度によるコントロールがしやすい音声合成の実現が期待できる

４．研究成果

(1) Simple recurrent unit を用いた DGP 音声合成の時系列拡張

音声は音響特徴量が連続的に変化することから，音声合成ではリカレント構造が広く使用される．代表的なリカレント構造としては LSTM や GRU といった構造がしばしば使用されるが，DGP のガウス過程回帰を直接に LSTM 構造に組み込むのは，計算量の問題から現実的でない．そこで，GPU による並列計算とシンプルな時系列方向の演算を組み合わせた，simple recurrent unit を DGP に組み込んだ．具体的には，右図のように並列計算部分の関数をガウス過程に従う関数とすることで，DGP におけるリカレント構造を実現した．

主観評価実験により提案法の合成音声はリカレント構造を持つ DNN 音声合成や，リカレント構造を持たない DGP 音声合成よりも自然性の高い音声を生成できることを示した．

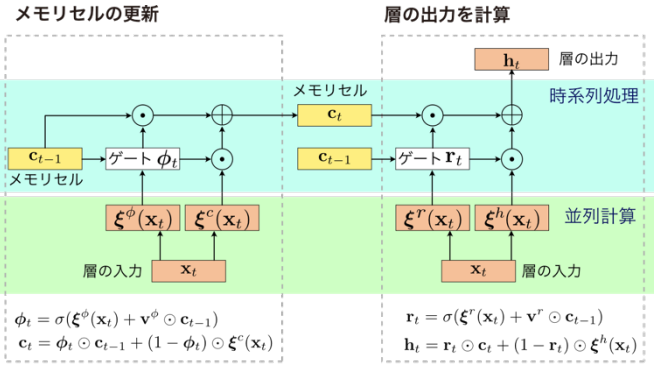

図 1 DGP 音声合成における Simple recurrent unit の構造

(2) DGP 音声合成の sequenece-to-sequence 型音声合成への拡張

近年の音声合成分野では，文字や音素などの単純な記号列から音声や音響特徴量の列を直接予測する sequence-to-sequence 型の音声合成が広く研究されており，DGP 音声合成においても sequence-to-sequence 型のモデルを導入することによって自然性の向上を目指した．

Sequence-to-sequence 型の音声合成では，長さの異なる記号列と音響特徴量列の長さを揃える length regulator が広く使用されており，本研究では DGP を用いた length regulator を提案した．また，必ずしも特徴が連続的に変化しない記号列のモデル化に有効な Self-attention 構造に対しても DGP に基づく構造を提案した．具体的には simple recurrent unit と同様に，DNN のパーセプトロンで表現される関数をガウス過程に基づく関数で置き換えた．

Model	MCD	MOS
SRU-DGP	5.87	3.41
Seq2Seq-SRU-NN	6.29	2.08
Seq2Seq-SA-NN	6.03	2.83
Seq2Seq-SRU-DGP	6.15	2.71
Seq2Seq-SA-DGP	6.01	2.96
ORIG	-	4.42

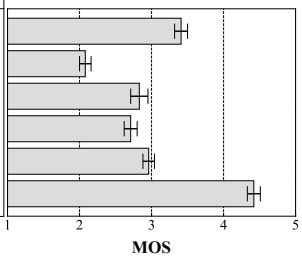

図 2 Sequence-to-sequence DGP 音声合成の mean opinion score テストの結果

音声合成実験の結果，提案法(Seq2Seq-SA-DGP)は同様の構造を持つ DNN である FastSpeech 型の従来法(Seq2seq-SA-NN)より自然な音声を生成できることを示した．以上により，DGP の拡張可能性として，リカレント型，self-attention，sequence-to-sequence といった，DNN のようなモデルの拡張は DGP でも可能であることを示した．

(3) DGP 潜在変数モデルを用いた多話者音声合成

DGP 音声合成の特長の一つとして，未知の入力の分布をベクトルの分布で表す潜在変数を自然に導入できることである．本研究では，複数の話者の音声を一つのモデルで表現する多話者音声合成において，話者の潜在変数を導入した DGP 潜在変数に基づく音声合成を提案した．モデル構造を各層の入力に潜在変数を結合する形で表現し，各話者の潜在変数を表すベクトルの事後分布を変分ベイズの枠組みにより推定した．

実験では，すべての話者を one-hot 表現する DGP 音声合成との比較を行い，特に目的話者の発話数が 5 文程度と少ない場合において，潜在変数モデルによる話者空間の推論が有効であることを示した．また，DNN で話者空間を表現する変分オートエンコーダに基づく多話者音声合成との比較を行い，DGP 音声合成の方が安定して高品質な音声を生成できることを示した．さらに，図に示す話者空間の評価により，学習データの話者をバランスよく分散させられることに成功し，生成された話者空間からランダム生成した場合でも自然な音声を生成できることを示した．

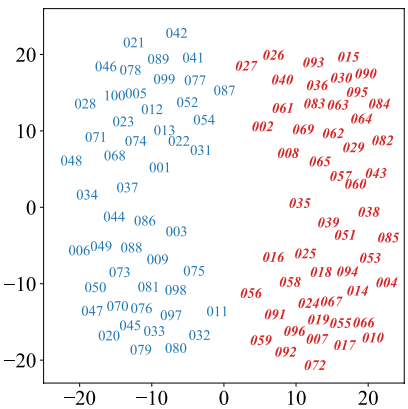

図 3 DGP 潜在変数モデルによる話者空間

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Mitsui Kentaro、Koriyama Tomoki、Saruwatari Hiroshi	４．巻 132
２．論文標題 Deep Gaussian process based multi-speaker speech synthesis with latent speaker representation	５．発行年 2021年
３．雑誌名 Speech Communication	６．最初と最後の頁 132～145
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.specom.2021.07.001	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計19件（うち招待講演　0件／うち国際学会　5件）

１．発表者名 Tomoki Koriyama, Hiroshi Saruwatari
２．発表標題 Utterance-level Sequential Modeling For Deep Gaussian Process Based Speech Synthesis Using Simple Recurrent Unit
３．学会等名 Proc. 45th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)（国際学会）
４．発表年 2020年

１．発表者名 Kentaro Mitsui, Tomoki Koriyama, Hiroshi Saruwatari
２．発表標題 Multi-speaker Text-to-speech Synthesis Using Deep Gaussian Processes
３．学会等名 Proc. Interspeech 2020（国際学会）
４．発表年 2020年

１．発表者名 Taiki Nakamura, Tomoki Koriyama, Hiroshi Saruwatari
２．発表標題 Sequence-to-Sequence Learning for Deep Gaussian Process Based Speech Synthesis Using Self-Attention GP Layer
３．学会等名 Proc. Interspeech 2021（国際学会）
４．発表年 2021年

１．発表者名
Kazuki Mizuta, Tomoki Koriyama, Hiroshi Saruwatari

２．発表標題
Harmonic WaveGAN: GAN-Based Speech Waveform Generation Model with Harmonic Structure Discriminator

３．学会等名
Proc. Interspeech 2021（国際学会）

４．発表年
2021年

１．発表者名
Kazuya Yufune, Tomoki Koriyama, Shinnosuke Takamichi, Hiroshi Saruwatari

２．発表標題
Accent Modeling of Low-Resourced Dialect in Pitch Accent Language Using Variational Autoencoder

３．学会等名
Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)（国際学会）

４．発表年
2021年

１．発表者名
三井健太郎，郡山知樹，猿渡洋

２．発表標題
多話者音声合成における深層ガウス過程潜在変数モデルを用いた音響モデル・話者表現の同時学習

３．学会等名
日本音響学会2020年秋季研究発表会講演論文集

４．発表年
2020年

１．発表者名
湯舟航耶，郡山知樹，猿渡洋

２．発表標題
変分オートエンコーダを用いたアクセントの潜在変数表現の検討

３．学会等名
日本音響学会2020年秋季研究発表会講演論文集

４．発表年
2020年

１．発表者名
中村泰貴，郡山知樹，猿渡洋

２．発表標題
深層ガウス過程音声合成におけるsequence-to-sequence学習の初期検討

３．学会等名
日本音響学会2020年秋季研究発表会講演論文集

４．発表年
2020年

１．発表者名
郡山知樹，猿渡洋

２．発表標題
活性化関数とカーネル関数の関係性を用いたガウス過程音声合成の評価

３．学会等名
日本音響学会2021年春季研究発表会講演論文集

４．発表年
2021年

１．発表者名
中村泰貴，郡山知樹，猿渡洋

２．発表標題
深層ガウス過程を用いたsequence-to-sequence音声合成のモデル構造の評価

３．学会等名
日本音響学会2021年春季研究発表会講演論文集

４．発表年
2021年

１．発表者名
水田和輝，郡山知樹，猿渡洋

２．発表標題
音声の周波数特性を考慮した畳み込み層を持つ波形生成モデルの検討

３．学会等名
日本音響学会2021年春季研究発表会講演論文集

４．発表年
2021年

１．発表者名
郡山知樹，猿渡洋

２．発表標題
深層ガウス過程音声合成における関数の確率微分方程式表現の利用の検討

３．学会等名
日本音響学会2020年春季研究発表会講演論文集，2-Q-44，pp.1127-1128．(Mar．2020)

４．発表年
2020年

１．発表者名
芹川武尊，郡山知樹，猿渡洋

２．発表標題
Attentionに基づく音声変換のためのアラインメント予測モデルの検討

３．学会等名
日本音響学会2020年春季研究発表会講演論文集，2-2-2，pp.1077-1078．(Mar．2020)

４．発表年
2020年

１．発表者名
三井健太郎，郡山知樹，猿渡洋

２．発表標題
深層ガウス過程に基づく多話者音声合成

３．学会等名
日本音響学会2020年春季研究発表会講演論文集，1-2-2，pp.1043-1044．(Mar．2020)

４．発表年
2020年

１．発表者名
郡山知樹，猿渡洋

２．発表標題
深層ガウス過程に基づく音声合成におけるリカレント構造を用いた系列モデリングの検討

３．学会等名
日本音響学会2019年秋季研究発表会講演論文集，1-P-25，pp.1025-1026．(Sept．2019)

４．発表年
2019年

１．発表者名
三井健太郎，郡山知樹，猿渡洋

２．発表標題
深層ガウス過程とアクセントの潜在変数表現に基づく音声合成の検討

３．学会等名
電子情報通信学会技術研究報告，vol.119，no.398，SP2019-49，pp.31-36．(Jan．2020)

４．発表年
2020年

１．発表者名
中村泰貴，郡山知樹，猿渡洋

２．発表標題
Self-Attention構造を有する深層ガウス過程を用いたSequence-to-Sequence音声合成

３．学会等名
日本音響学会2021年秋季研究発表会講演論文集

４．発表年
2021年

１．発表者名
湯舟航耶，郡山知樹，高道慎之介，猿渡洋

２．発表標題
VQ-VAEに基づくアクセントの潜在変数表現を用いた方言音声合成

３．学会等名
日本音響学会2021年秋季研究発表会講演論文集

４．発表年
2021年

１．発表者名
湯舟航耶，郡山知樹，高道慎之介，猿渡洋

２．発表標題
アクセント潜在変数を用いた方言音声合成における文単位生成の評価

３．学会等名
電子情報通信学会技術研究報告

４．発表年
2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関