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　２ 年　 ６ 月　 ８ 日現在

機関番号：１２６０１
研究種目：若手研究(B)
研究期間：2017～2019
課題番号：１７Ｋ１２７１１

研究課題名（和文）ガウス過程潜在変数モデルを用いた韻律の分散表現

研究課題名（英文）A Study on Prosody Embedding Based on Gaussain Proceess Latent Variable Model

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・助教

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）：統計的音声合成において，音声合成におけるラベルはテキストだけではなく，テキストに含まれない韻律情報などを含める必要がある．そこで，アクセントなどの韻律情報を音声から自動獲得する手法として，ガウス過程潜在変数モデルを用いた，韻律の分散表現の開拓・確立を行った．本研究ではまず複雑な言語特徴量から特徴抽出を行うモデルとして，深層ガウス過程による音声合成手法の確立を行った．深層ガウス過程では未知の韻律情報を確率変数として推定可能なため，ラベルの付与されていないデータに対して韻律の潜在変数を推論することで，音声合成の半教師あり学習の有効性を示した．

研究成果の学術的意義や社会的意義
音声合成におけるラベルはテキストだけではなく，テキストに含まれない韻律情報などを含める必要があり，話し言葉やオーディオブックなどの多様な音声合成システムを構築する際には，ラベルを付与に係るコストなどの問題が生じる．また，同じテキストであっても文脈によって読み方が変わることによるテキストからの韻律推定の困難さや，ラベリングを行う人物間でのラベルの不一致が生じる．そこで本研究では機械学習により韻律を低次元の潜在空間で表現する自動化手法を提案し，データベース構築の容易さや，多様な韻律表現による豊かな音声合成の構築への基礎の構築を行った．

研究成果の概要（英文）：In statistical speech synthesis, labels in speech synthesis must include not only text but also prosodic information. As a method to obtain latent prosodic information such as accent from speech, we proposed speech synthesis using Gaussian process latent variable model. In this study, we first investigate a speech synthesis system based on deep Gaussian processes, which can extract hidden embedding from complicated language features. The speech synthesis can infer unknown prosodic information as a randam variable of probabilistic model. Therefore, we proposed a semi-supervised speech synthesis system, in which labeled and unlabeled speech data is used as a trainind data by estimating latent prosodic features of the unlabeled speech data.

研究分野：音声情報処理

キーワード：音声情報処理　韻律　ガウス過程　機械学習　音声合成

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

テキスト音声合成技術の発展に伴い，音声合成の利用場面はスマートフォン，コミュニケーションロボット，映像コンテンツなどへの広がりを見せており，東京オリンピックにおける多言語翻訳技術の実現など，社会における音声情報処理への需要が高まっている．一方で近年の国外の研究動向として，画像処理や音声認識で注目を集めたディープニューラルネットワークを音声合成に応用した技術が広く研究され，実用化に向け技術が確立しつつある．そのような背景のもと，音声合成技術のコンペティションである Blizzard Challenge では絵本の読み聞かせがタスクとして選ばれるなど，多様な用途への期待に応えるための取り組みが広がっている．
ここで，問題となるのが学習データのラベリングである．一般的に音声とその発話内容を記述したラベルが大量に用意できれば，高品質な音声を合成することが可能である．しかし，音声合成におけるラベルはテキストだけではなく，テキストに含まれない韻律情報などを含める必要があり，話し言葉やオーディオブックなどの音声合成システムを構築する際には以下のような問題が生じる．
1.音声データに韻律などのラベルを付与するにはコストを要する．
2.同じテキストであっても文脈によって読み方が変わる．そのためテキストからの韻律の推定は困難である．
3.ラベルを付与する場合においても．ラベリングを行う人物間でのラベルの不一致が生じる．
このような問題を解決するためには，韻律情報ラベルを自動的に学習データに付与し，合成時にも自動的にそのラベルを付与できるシステムの実現が望まれる．

２．研究の目的

先行研究として，タイ語音声合成において，ガウス過程潜在変数モデルに基づく基本周波数曲線の低次元表現がある．この研究では，音声の基本周波数の系列を低次元の連続空間へマッピングし，テキストのみから推定することが困難なストレスの有無の分離を実現している．
本研究ではさらにこれを発展させ，より広範囲の韻律情報を自動的に推論可能な音声合成システムを構築することである．そのためには，具体的にはテキストのみから決定することが困難な日本語のアクセントを対象として，日本語のアクセントの低次元表現によって，韻律ラベルの付与されていない音声が大量に含まれる場合であっても学習可能な音声合成システムの構築を目指す．

3. 研究の方法

(1) ガウス過程の深層モデルを用いた音声合成の提案

当初の計画では，従来のガウス過程潜在変数モデルに基づく基本周波数曲線の低次元表現を，日本語のアクセントに適用する予定であった．しかしながら，単純なガウス過程潜在変数モデルでは，音声から言語情報と潜在的な韻律情報を分離することは困難であった．そこで，本研究課題ではまず，複雑な言語特徴量を深層ベイズモデルによって扱いやすい形にする手法として，深層ニューラルネットワーク(deep neural network，DNN)と深層ガウス過程(deep Gaussian process，DGP)を用いた音声合成手法の検討を行う．

(2) 深層モデルとガウス過程潜在変数モデルを組み合わせた韻律の潜在表現の獲得

上記の深層モデルをガウス過程潜在変数モデルに拡張し，複雑な言語特徴量を分離しつつ，潜在的な韻律情報を分離する．このとき，少量の韻律ラベルありデータと大部分の韻律ラベルなしデータから半教師あり学習により，韻律の潜在空間と音声合成システムの同時学習を行い，少量の韻律ラベルありのデータのみを用いた場合および韻律ラベルを使用しない場合との比較を行う．

(3) 多様な音声合成における深層ガウス過程の応用

深層ガウス過程(DGP)は，DNN と比較して，カーネル回帰によるより柔軟なモデル表現が可能である，ベイズ学習によるモデルの複雑さを考慮した最適化が可能であるといったメリットがあり，(1)の発展として DGP を用いた音声合成の拡張可能性を検討する．具体的には多話者を同時にモデル化する多話者音声合成に深層ガウス過程を用いたときの性能を DNN 多話者音声合成と比較する．

４．研究成果

(1)ガウス過程の深層モデルを用いた音声合成

本研究課題ではまず，既存のガウス過程回帰(GPR)に基づく音声合成を深層学習に拡張する手法を提案した．具体的には GPR 音声合成の特徴抽出として DNN を用いる GP-DNN と，すべての層に GPR を用いる DGP (DeepGP)を比較した．
データベースには音声合成システム XIMERA 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 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，およびそれらの動的特徴量と有声／無声フラグを音響特徴量として使用した．また，継続長モデルでは音素継続長を音響特徴量とした．

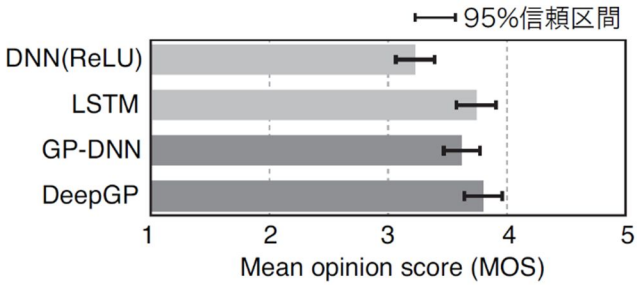

図 1 GP-DNN と DeepGP の主観評価結果

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．

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．結果を図 1 に示す．DNN 音声合成と他手法を比較すると，α=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 DGP 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．

(2)深層ガウス過程の潜在変数モデルを用いた音声合成

複数のガウス過程による階層モデルである深層ガウス過程(DGP)を使用し，図 2 に示す韻律の潜在変数モデルを提案した．DGP は回帰モデルとしてだけでなく潜在変数モデルとして使用することも可能であり，1 層の GP の枠組みにおいて，入力変数が部分的に欠損しているときに，欠損部分を潜在変数で表現し半教師あり学習を行う semi-described GP と呼ばれる手法が既に示されている．

そこで本研究では， semi-described GP と DGP を組み合わせた半教師あり学習による音声合成の有効性を検討した．具体的には，少量のアクセントラベル付きデータとアクセントラベルなしデータを使用し，コンテキストのうちアクセント型に関係のない部分を観測データと， 関係する部分を欠損値とそれぞれ見なし，アクセント情報を潜在変数で表現する．このとき，1 層の GP ではなく DGP を用いることで 潜在変数表現からそのまま音声合成のモデルの学習ができるため，韻律情報を明示的に予測する必要がないという特長がある．

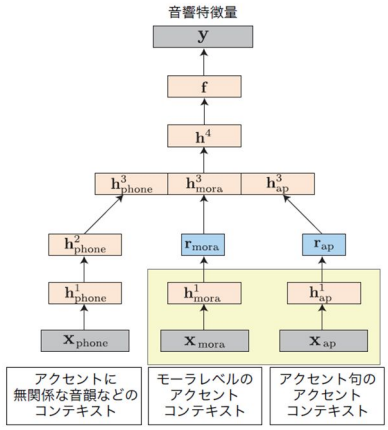

図 2 アクセントの潜在変数表現を用いた DGP のネットワーク構造

実験条件として，データベースには日本語音声合成システム XIMERA 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． さらに，学習データの約 9 割にあたる 1434 文をアクセントラベルなしのデータとし， 残りの 159 文のうち 99 文をラベル付きデータとして，60 文を開発セットとしてそれぞれ使用した．また，入力のコンテキストベクトルとして，基本情報，モーラアクセント情報，句アクセント情報はそれぞれ， 477，38，99 次元とした．評価として，潜在変数表現を使用しない DGP モデルとの比較を行った． その際，全データに対しアクセントラベルが付与されている理想的な場合(FULL)と， 全データを使用するがアクセント型を学習に使用しない場合(W/O ACCENT)， アクセントラベルの付与されている 99 文だけを用いる場合(LABELED)を条件として用いた．

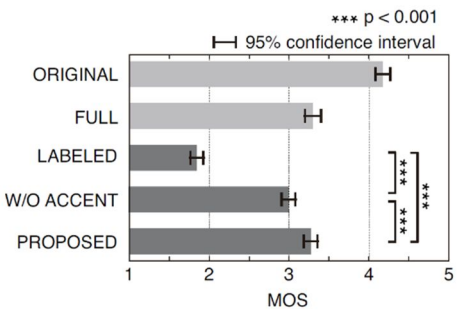

図 3 韻律の潜在変数モデルの主観評価結果

合成音声の聴覚上の評価を行うため主観評価実験を行った． 聴取者は 45 名で，各聴取者はランダムに選ばれた 4 文に対して，FULL，PROPOSED，W/O ACCENT，LABELED の各手法で合成した音声と原音声(ORIGINAL)の自然性を 5 段階のスコア(1：非常に悪い，2：悪い，3：ふつう，4：良い，

5：非常に良い)で評価した．結果の Mean opinion score (MOS) 値を図 3 に示す． 結果から理想的な条件で合成した FULL と 99 文のみに韻律ラベルの付与されている PROPOSED ではほとんど差が見られなかった．したがって，提案法の潜在変数モデルによって学習データの 10%程度にしかラベルが付与されていない場合でも，十分に自然な音声が合成可能であることを示した．

最後に提案法によって合成した音声の基本周波数の例を図 4 に示す．韻律ラベルが付与されているデータのみを用いたためデータ量の少ない LABELED では F0 曲線が平坦になってしまっている．また，韻律ラベルを使用しない W/O ACCENT では曲線のピークが理想的な条件の FULL より低く，アクセントが知覚されにくい曲線となっている．それらに対し，提案法の PROPOSED では FULL に近い F0 曲線が得られ，この結果が主観評価結果に反映されていると考えられる．

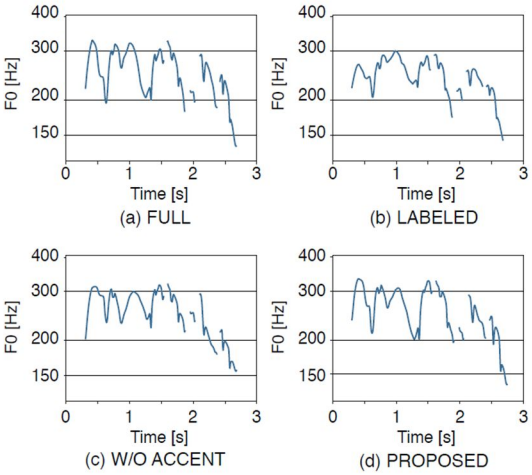

図 4　提案法による基本周波数(F0)曲線

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 Tomoki Koriyama, Takao Kobayashi	４．巻 vol.27, no.5
２．論文標題 Statistical Parametric Speech Synthesis Using Deep Gaussian Processes	５．発行年 2019年
３．雑誌名 IEEE/ACM Transactions on Audio, Speech, and Language Processing	６．最初と最後の頁 948-959
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TASLP.2019.2905167	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12件（うち招待講演　0件／うち国際学会　3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), (May 2020)（国際学会）
４．発表年 2020年

１．発表者名 Tomoki Koriyama, Takao Kobayashi
２．発表標題 Semi-Supervised Prosody Modeling Using Deep Gaussian Process Latent Variable Model
３．学会等名 Proc. 20th Annual Conference of the International Speech Communication (INTERSPEECH 2019), pp.4450-4454. (Sept. 2019)（国際学会）
４．発表年 2019年

１．発表者名 Tomoki Koriyama, Takao Kobayashi
２．発表標題 A Training Method Using DNN-guided Layerwise Pretraining For Deep Gaussian Processes
３．学会等名 Proc. 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019), pp.4785-4789. (May 2019)（国際学会）
４．発表年 2019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

４．発表年
2020年

１．発表者名
郡山知樹，小林隆夫

２．発表標題
深層ガウス過程に基づく音声合成のための事前学習の検討

３．学会等名
日本音響学会2018年秋季研究発表会講演論文集

４．発表年
2018年

１．発表者名
郡山知樹，小林隆夫

２．発表標題
GPR音声合成のための深層構造の利用の検討

３．学会等名
日本音響学会2018年春季研究発表会講演論文集，pp．1507-1508

４．発表年
2018年

１．発表者名
郡山知樹，小林隆夫

２．発表標題
GPR音声合成における深層ガウス過程の利用の検討

３．学会等名
電子情報通信学会技術研究報告，Vol．117，No．517，pp．27-32

４．発表年
2018年

１．発表者名 郡山知樹，小林隆夫
２．発表標題 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
３．学会等名 電子情報通信学会技術研究報告，Vol．117，No．393，pp．5-10
４．発表年 2018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