論文 / 著書情報
Article / Book Information

論題(和文)	ガウス過程回帰に基づく音声合成システムの評価
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2015年秋季研究発表会講演論文集, Vol. , No. , pp. 235-236
Citation(English)	, Vol. , No. , pp. 235-236
発行日 / Pub. date	2015, 9

ガウス過程回帰に基づく音声合成システムの評価*

◎郡山知樹, 小林隆夫 (東工大)

1　はじめに

我々はこれまでに，ガウス過程回帰 (GPR) に基づくフレームレベル音響モデリングを用いた統計的パラメトリック音声合成の枠組み (GPR 音声合成)[1] を提案しており，従来の HMM 音声合成 [2] に比べ高い自然性を得られることを報告している．一方で，近年ディープニューラルネット (DNN) を用いた音声合成システムも提案され [3]，その性能について文献 [4] で詳細な検討が行われている．DNN の深層構造の持つ柔軟性から DNN 音声合成もまた，自然性の向上に有効であることが示されている．本稿では，音声合成における GPR の有効性の評価のため，DNN 音声合成とのシステムおよび合成音声の比較評価を行う．

2　音声合成システムの比較

本稿ではまず，HMM，DNN および GPR に基づく音声合成システムの共通部分を考える．これらのシステムでは音声波形から得られたメルケプストラムや LSP，対数 F0 といった音響特徴量とテキストから得られるコンテキストとの関係をなんらかの統計モデルで表現している．また，合成時にはモデルにより得られた音声パラメータ系列だけでなく，動的特徴量や GV などのポストフィルタを使用することで，滑らかで自然性な音声を生成する．

一方で，これらの手法の差異は Fig. 1 に示すように，音声パラメータ系列の予測に顕著に表れる．HMM 音声合成における予測は決定木を用いて行う．決定木のノードに音声パラメータの確率分布がそれぞれ対応しており，予測時にはコンテキストに対応する確率分布を HMM の状態レベルで決定し，その分布を用いて音声パラメータ系列の生成を行う．

DNN 音声合成では決定木の代わりにフィードフォワードのニューラルネットワークを用いている．複数の層で構成されるニューラルネットは決定木に比べ複雑な構造であることから，コンテキスト同士の依存関係なども表現可能になっていると考えられる．また，DNN 音声合成ではフレームレベルのコンテキストからフレームレベルの音声パラメータを予測するネットワークを構築する．そのため，HMM 音声合成における状態などのセグメントレベルでの予測に比べ，より細かい音声パラメータの変化をモデル化可能である．

GPR 音声合成はクラスタの決定と，ガウス過程に基づく回帰の 2 つの処理により予測を行う．クラスタの決定は決定木を用いて行い，合成したい音声パラメータ列と類似した事例を持つフレームを得る．次に，合成したいフレーム列と局所的な特徴を持つクラスタ内のフレーム，それに加えグローバルな特徴を持つ補助点とのグラム行列を計算する．このグラム行列はフレームレベルの音響特徴量の同時ガウス分布における共分散行列であり，その値はフレーム間のコンテキストの類似度を表すカーネル関数で与えられる．

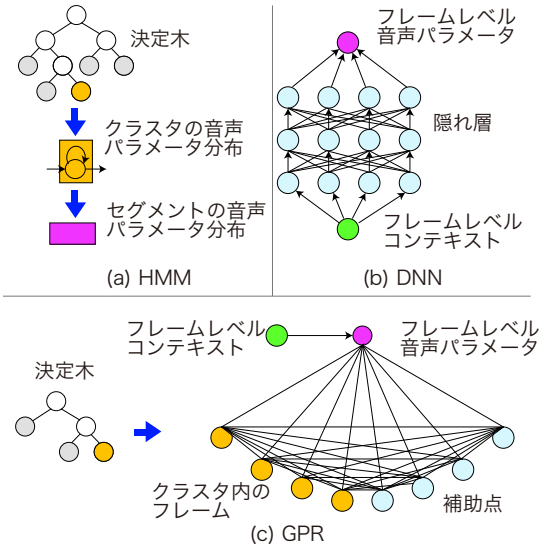

Fig. 1　HMM，DNN，GPR 音声合成の音響特徴量予測手法の比較

GPR 音声合成を HMM 音声合成と比較すると，決定木の利用という面では共通である．しかし，HMM 音声合成では未知のコンテキストへの対応のためにモデルの記述長を基準として木が構築されるが，GPR 音声合成では計算量の削減のためにクラスタ内のフレーム数を基準として木を構築するため，結果として得られる決定木は異なる構造となる．また，GPR 音声合成と DNN 音声合成を比較すると，両手法は共にフレームレベルの回帰を行うことで音声パラメータの予測を行う．しかし，ネットワーク構造は大きく異なり，DNN は深い有向グラフで表されるのに対し，GPR は Fig. 1(c) のように浅い密なネットワークで表される．加えて，DNN が音響特徴量を隠れ層のモデルパラメータのみで表すのに対し，GPR はクラスタ内の事例をそのまま利用するノンパラメトリックモデルとなるのが両手法の相違点である．

3　実験

3.1　実験条件

実験には ATR 日本語音声データベースセット B に含まれる女声話者，男声話者，各 2 名の計 4 名 (FKS, FTK, MMY, MHT) の音声を用いた．学習データには 450 文を，テストデータには学習データに含まれない 53 文を用いた．周波数 16kHz でサンプリングされた音声に対し，STRAIGHT を用いて基本周波数およびスペクトル包絡，非周期性指標を抽出した．スペクトル特徴量には STRAIGHT スペクトルから得られる 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 Δ，Δ^2 の動的特徴量を音響特徴量として使用した．

HMM 音声合成に使用するモデルは 5 状態の left-to-right スキップなし隠れセミマルコフモデル (HSMM)

*Evaluation of speech synthesis system based on Gaussian process regression. by KORIYAMA, Tomoki and KOBAYASHI, Takao (Tokyo Institute of Technology)

Table 1　客観評価結果．値はそれぞれ，音素継続長の RMS 誤差 [ms]，メルケプストラム距離 [dB]，非周期性指標のスペクトル距離 [dB]，有声／無声推定精度 (F 値)，対数 F0 の RMS 誤差 [cent] の全話者の平均を表す

	音素継続長	メルケプストラム	非周期性指標	有声／無声フラグ	対数 F0
HMM	22.5	4.86	2.85	0.957	205
DNN 3-layer	19.7	4.68	2.74	0.966	178
DNN 6-layer	20.0	4.65	2.72	0.965	184
GPR State	21.8	4.66	2.78	**0.969**	174
GPR Phone	**19.3**	4.54	2.76	0.967	**169**
GPR Mora	27.9	**4.37**	**2.59**	0.968	170
GPR Phrase	29.3	4.83	2.78	0.964	185

とし HSMM の各状態の出力分布は対角共分散行列を持つ単一ガウス分布とした．音素，モーラ，アクセント句，呼気段落，文長に関するコンテキストを用いて，状態単位のクラスタリングを行った．その際，最小記述長 (MDL) を決定木の停止基準とした．

DNN 音声合成の入力変数には，HMM 音声合成のコンテキストに対する質問セットから得られる 450 個の 2 値変数と GPR 音声合成で使用する 46 個の位置情報を使用した．入力変数は $[0.01, 0.99]$ の範囲に正規化し，出力変数は平均 0，分散 1 となるように正規化を行った．文献 [4] の結果を参考に，ニューラルネットの隠れ層には 3 層ないし 6 層を使用し，各層の素子数は 1024，アクティベーション関数は tanh とした．パラメータ推定時のミニバッチのサイズは 256 とし，AdaGrad[5] を用いて学習率の自動決定を行った．DNN の学習時には学習データに含まれる 50 文章を開発セットとし，開発セットの誤差によって学習の停止基準を決定した．なお，DNN の学習および予測の実装には Pylearn2[6] を使用した．

ガウス過程回帰・分類に用いる PIC 近似 [7] におけるブロックの最大フレーム数は 1024 とし，学習データに含まれるフレーム全体からランダムに選択した 1024 フレームを疑似データセットの補助点とした．このとき，PIC 近似におけるブロックの決定には HMM 音声合成の枠組みで得られるコンテキスト決定木を使用した．ただし，本研究ではクラスタリングの単位として状態 (State)，音素 (Phone)，モーラ (Mora)，アクセント句 (Phrase) を導入し比較を行った．モーラおよびアクセント句単位でのクラスタリングにはそれぞれモーラ，アクセント句単位の HSMM を学習し決定木の構築を行った．

3.2　結果

合成音声の原音声に対する音響特徴量の歪および有声／無声フラグの推定精度の 4 話者の平均値を Table 1 に示す．DNN の 3-layer および 6-layer は隠れ層の数を，GPR の State, Phone, Mora, Phrase はそれぞれクラスタリングの単位を表す．GPR におけるクラスタリングの単位を比較すると，音素継続長においては音素がメルケプストラムおよび非周期性指標においてはモーラがそれぞれ他の単位に比べ歪が大きく減少している．また GPR と DNN を比較すると，クラスタリング単位にモーラを用いた GPR は，DNN に比べスペクトル歪や対数 F0 歪が小さくなっていることがわかる．このような結果を得た理由として，GPR 音声合成が事例をそのまま利用するノンパラメトリックであること，また，モーラ単位のクラスタリングによりひとまとまりの連続的な事例を得ていること，が考えられる．

次に，合成音声の自然性を対比較実験により比較した．DNN の隠れ層は 3 層とし，GPR で行うクラスタリングにおいて，音素継続長の単位を音素，継続長以外の特徴量の単位をモーラとした．主観評価実験における被験者は 6 人で，各被験者に対し 10 文章を話者ごとにランダムに選択した．本実験では客観評価結果から被験者は差がないと感じた場合 Neutral を選べるものとした．結果を Table 2 に示す．表から HMM に対し DNN，GPR は有意にスコアが高いことがわかる．また，DNN と GPR を比較すると，Neutral と判断された合成音声が多く，有意な差は見られなかったものの，GPR 音声合成は DNN 音声合成に比べわずかに高いスコアを得た．

Table 2　対比較テストによる主観評価結果 [%]

HMM	DNN	GPR	Neutral	p 値	Z 値
15.5	**55.7**		28.9	$< 10^{-10}$	-12.0
14.0		**52.4**	33.6	$< 10^{-10}$	-11.6
	19.9	23.5	56.5	0.261	1.1

4　おわりに

本稿では，ガウス過程回帰 (GPR) を用いた音声合成システムと HMM および DNN 音声合成との比較評価を行った．客観および主観評価の結果 GPR 音声合成は DNN 音声合成に比べ，音響特徴量歪が小さく，僅かな差ではあるが自然性が高いという評価結果を得た．

謝辞　本研究の一部は，JSPS 科研費 15H02724 の助成を得た．

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検討，” 音講論 (春)，2-2-9, pp.269–270，2015.

[2] 吉村 他，“HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化，” 信学論，J83-D-II(11), 2099–2107, 2000.

[3] Zen et al., “Statistical parametric speech synthesis using deep neural networks,” in Proc. ICASSP, 2013, pp.7962–7966.

[4] Qian et al., “On the training aspects of deep neural network (DNN) for parametric TTS synthesis,” in Proc. ICASSP, 2014, pp.3829-3833.

[5] Duchi et al., “Adaptive subgradient methods for online learning and stochastic optimization,” The Journal of Machine Learning Research, vol. 12, pp. 2121–2159, 2011.

[6] Pylearn2 http://deeplearning.net/software/pylearn2/.

[7] E. Snelson and Z. Ghahramani, “Local and global sparse Gaussian process approximations,” Proc. AISTATS, pp.524–531, 2007.