論文 / 著書情報
Article / Book Information

論題(和文)	GPR音声合成のためのフレームコンテキストカーネルに基づく決定木構築の検討
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2017年秋季研究発表会講演論文集, Vol. , No. , pp. 177-178
Citation(English)	, Vol. , No. , pp. 177-178
発行日 / Pub. date	2017, 9

GPR 音声合成のためのフレームコンテキストカーネルに基づく決定木構築の検討 *

○郡山知樹, 小林隆夫 (東工大)

1 はじめに

我々はこれまでに新たな統計的音声合成手法の枠組みとしてガウス過程回帰 (Gaussian Process Regression: GPR) に基づく音声合成手法を提案している [1]. GPR 音声合成は，フレームレベルのコンテキストの類似度を表すカーネル関数を用いるノンパラメトリックモデルによって，学習データの音響特徴量系列から直接合成音声を音響特徴量を予測し，自然な音声を生成することを目的としている．このとき，単純な GPR の計算量は学習データのフレーム数の 3 乗のオーダーとなってしまうため，GPR 音声合成では学習データをいくつかのブロックに分割し計算量を削減する部分独立条件 (PIC) 近似 [2] を用いているが，ブロックの分割方法についての検討は十分に行われていない．

これまでの研究 [1] では隠れマルコフモデル (HMM) によりセグメントをモデル化し，HMM の尤度に基づき決定木クラスタリングを行うことでブロック分割を行うハイブリッド手法を用いていたが，HMM の尤度基準が必ずしも GPR に効果的であるという合理性はない．さらに HMM を用いるシステムでは，音素やモーラなどのクラスタリング単位をヒューリスティックに選択する必要があるだけでなく，主に 2 値で表現される HMM のためのコンテキストと，主に連続値で表される GPR のためのコンテキストを共に定義する必要があった．そこで本研究では新たなブロック分割法として，HMM を用いず GPR のためのフレームレベルコンテキストのみを用いたフレーム単位での決定木構築手法を提案する．

2 カーネル PCA

本研究では決定木の構築にフレームレベルのコンテキストに対するカーネル関数の出力を用いる．ただし，カーネル関数は 2 入力関数であり，そのままでは扱いが容易ではない．そこで本研究では，カーネル主成分分析 (PCA)[3] を用いて入力変数の低次元空間へ射影したベクトルを使用する．

カーネル PCA はグラム行列 $\mathbf{K} = K(\mathbf{X}, \mathbf{X})$ の行および列方向の平均を 0 にした中心化グラム行列 $\tilde{\mathbf{K}}$ に対して，$\tilde{\mathbf{K}}$ の固有値問題を解くことに帰着される．カーネル PCA を用いて p 次元空間に射影するとき，固有値を要素に持つ対角行列 $\mathbf{\Lambda} = \mathrm{diag}[\lambda_1, \ldots, \lambda_p]$ および固有ベクトルを要素に持つ $N \times p$ 行列 $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_N]^\top$ を用いて，第 i フレームの特徴ベクトルは

$$\boldsymbol{\phi}(\mathbf{x}_i) = \mathbf{\Lambda}^{1/2}\mathbf{v}_i \qquad (1)$$

で表される．また未知の入力 $\mathbf{x}'$ を射影した特徴ベクトルは，中心化したグラム行列を求める関数 $\tilde{K}$ を用いて，次の式で求められる．

$$\boldsymbol{\phi}(\mathbf{x}') = \left(\tilde{K}(\mathbf{x}', \mathbf{X})\mathbf{V}\mathbf{\Lambda}^{-1/2}\right)^\top \qquad (2)$$

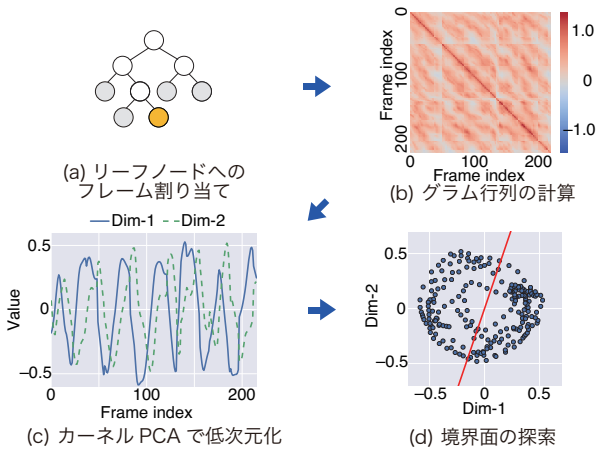

Fig. 1 決定木構築におけるリーフノード分割のアウトライン

3 決定木構築手法

カーネル関数に基づいた分割によって GPR に適した決定木を構築する方法を考える．そのためにはまず，カーネル関数自体がデータに適合する必要がある．本研究では，近年提案された確率的変分 (SV)GP[4] に基づく，ハイパーパラメータ（カーネル関数パラメータおよび補助点）の最適化を行う．この手法では GP の周辺分布の変分下限をデータサンプルごとの値の和で近似するため，大量のデータに対してもミニバッチに基づく確率的勾配法を用いた最適化が可能である．

ハイパーパラメータを最適化した後，学習データの分割および合成文のコンテキストに対するブロックの割り当てを行う決定木を構築する．決定木の構築の手続きを以下に示す．

1. ルートノードのみの決定木を作る．
2. 学習データからランダムにフレームを選択し，決定木を用いてリーフノードに割り当てる．(図 1(a))
3. リーフノードのフレーム数がブロックの最大フレーム割り当て数以下であれば 2 に戻る．
4. カーネル PCA を用いてリーフノードに含まれるフレームの低次元の特徴ベクトルを求める．(図 1(b)(c))
5. 低次元空間を適切に分割する境界面を探索し，リーフノードを分割する．(図 1(d))
6. 3 に戻る．

境界面の探索には，コンテキストのみを用いる K-means 法に基づく分割と，コンテキストと音響特徴量の双方を用いる周辺尤度に基づく分割の 2 手法を

---

*A Study on Construction of Decision Tree Using Frame-level Context Kernel for GPR-based Speech Synthesis. by KORIYAMA, Tomoki, KOBAYASHI, Takao (Tokyo Institute of Technology)

提案する．K-means 法に基づく分割では低次元の特徴ベクトルに対して $K = 2$ の K-means クラスタリングで分割を行う．

周辺尤度に基づく分割では，分割したときの周辺尤度を最大にする第 2 種の最尤推定の枠組みで境界面を探索する．この手法では，コンテキストの類似性だけでなく，音響特徴量の類似性も考慮して分割を行うことができる．このとき任意の境界で周辺尤度を最大にしようとすると,「1 対その他のフレーム」のように不均衡な分割がしばしば発生してしまう．そこで，本研究ではカーネル PCA で得られる低次元空間において，図 1(d) のように原点を通る超平面を境界面として分割を行う．カーネル PCA で得られる特徴ベクトルの各次元における平均は 0 となるため，原点を通る境界面で分割した場合，フレーム数の偏りが小さくなることが期待できる．ここで，周辺尤度を最大化する境界面は解析的に求められないため，未知の関数に対する全体最適化手法である，ベイズ的最適化 [5] を用いて境界面の探索を行う．

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[6] 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 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 を音響特徴量として使用した．

コンテキストには音素，アクセント句などの開始，終了からの相対位置および，音素弁別特性や高低アクセントの情報からなる，579 次元の情報を用いた．GPR の部分独立条件 (PIC) 近似の補助点数は 1024，各ブロックの最大フレーム数は 1024 とした．SVGP に基づくハイパーパラメータ最適化においてはミニバッチサイズ 1024 で Adam に基づく最適化を行った．K-means および周辺尤度に基づく分割におけるカーネル PCA の次元は 10 とし，決定木の構築は各特徴量ごとに行った．従来法における HMM を用いた決定木構築ではクラスタリングの単位は音素とし，有声／無声フラグモデルの決定木にはメルケプストラムの木を用いた．

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表 1 に示す．表中の手法 SV はブロック分割を行わず確率的変分 (SV)GP の結果を直接音声合成に使用したものを表す．また，HMM-tree, K-means, Marginal はそれぞれ従来手法の HMM，提案手法の K-means 法および周辺尤度に基づく決定木構築を表す．表から補助点のみを用いる SV に比べ，ブロック分割を用いる他手法の歪が小さくなっていることがわかる．また，HMM-tree，K-means，Marginal では有声／無声フラグを除きほとんど歪に差が見られなかった．

主観評価実験では対比較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．結果を表 2 に示す．K-means と Marginal を比較すると K-means の方が有意に自然であるという結果を得た．また，HMM-tree と K-means との比較では K-means, HMM-tree と Marginal の比較では HMM-tree がそれぞれわずかにスコアが高いものの，有意な差は見られなかった．

Table 1　客観評価結果．MCEP: メルケプストラム距離 [dB]，f_o: 対数 f_o の RMS 誤差 [cent]，V/UV: 有声／無声誤り，BAP: 非周期性指標歪 [dB]，DUR: 音素継続長の RMS 誤差 [ms].

手法	MCEP	f_o	V/UV	BAP	DUR
SV	5.21	184	5.22	3.34	17.4
HMM-tree	5.12	178	5.08	3.28	16.4
K-MEANS	5.13	179	4.83	3.29	16.5
Marginal	5.14	179	4.91	3.29	16.6

Table 2　対比較による主観評価結果 [%]

HMM-tree	K-means	Marginal	Neutral	p
20.0	24.8		55.2	0.30
24.8		21.4	53.8	0.47
	26.7	14.3	59.0	< 0.005

5 おわりに

本稿では GPR 音声合成において，HMM を用いずにフレームレベルのコンテキストとカーネル関数だけで GPR における PIC 近似のためのブロック分割を行う手法を提案した．提案手法では，各リーフノードに対し カーネル PCA により得られた低次元空間を分割する境界面を決定する．主観評価の結果より，HMM を用いずに K-means を用いて境界面を決定する手法の有効性を示した．今後は，様々なデータを用いた詳細な検討や，低次元空間の次元や逐次的分割におけるデータ選択の順序が合成音声に与える影響の調査が必要である．

謝辞　本研究は JSPS 科研費 JP15H02724 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] E. Snelson & Z. Ghahramani, “Local and global sparse Gaussian process approximations,” Proc. AISTATS, pp.524–531, 2007.
[3] J. Taylor & M. Cristianini, *Kernel methods for pattern analysis,* Cambridge Univ. Press, 2004.
[4] J. Hensman et al., “Scalable Variational Gaussian Process Classification,” Proc. AISTATS, pp. 315–360, 2015.
[5] D. R. Jones et al., “Efficient global optimization of expensive black-box functions,” Journal of Global Optimization, 13(4), pp. 455-492, 1998.
[6] 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.