# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 論題(和文) | ガウス過程回帰に基づくF0パタン生成の検討 |
| Title(English) | |
| 著者(和文) | 郡山知樹, 能勢 隆, 小林隆夫 |
| Authors(English) | Tomoki Koriyama, Takashi Nose, Takao Kobayashi |
| 出典(和文) | 日本音響学会2014年秋期研究発表会講演論文集, Vol. , No. , pp. |
| Citation(English) | , Vol. , No. , pp. |
| 発行日 / Pub. date | 2014, 9 |

# ガウス過程回帰に基づくF0パタン生成の検討*

◎郡山 知樹 (東工大)，能勢 隆 (東北大)，小林 隆夫 (東工大)

## 1 はじめに

我々はこれまでに，統計的音声合成のためのガウス過程回帰 (GPR) に基づくフレームレベルのスペクトルモデルを提案し，自然性の高い音声が合成可能であることを示した [1]．これまでの報告では，フレームレベルのコンテキストとして音素特性とフレームの相対位置を用いることでスペクトルに関するフレームの相関関係を表していたが，これらのコンテキストは音素の類似性に関するものでありF0パタンのモデル化に直接用いることは困難である．そこで本稿では，アクセントなどの韻律情報に対してもフレームレベルのコンテキストを定義することで，GPRに基づくフレームレベルの音響モデリングの枠組みにおけるF0のモデル化手法を検討する．また，F0パタンの生成にはフレームの有声／無声を決定する必要があるため，本研究ではガウス過程分類に基づく有声／無声推定モデルを提案する．

## 2 フレームコンテキストの拡張

文献 [1] では，音素の基本的な特徴である音素弁別特性と音素内位置情報をフレームレベルのコンテキスト（フレームコンテキスト）を用いることで，スペクトル特徴量の変化がモデル化可能であることを示した．そこで，アクセント情報に対しても，アクセントの基本的な特徴であるモーラの高低，句頭のピッチ上昇（句頭音調），アクセント核をコンテキストに導入することで，F0パタンのモデル化を目指す．

本研究で提案するフレームコンテキスト $\mathbf{x}_n$ は

$$\mathbf{x}_n = (\mathbf{x}_{n,1}, \ldots, \mathbf{x}_{n,K}) \quad (1)$$

$$\mathbf{x}_{n,k} = (\mathbf{p}_{n,k}, \mathbf{c}_{n,k}) \quad (2)$$

$$\mathbf{p}_{n,k} = \left(\mathbf{p}_{n,k}^{(-1)}, \mathbf{p}_{n,k}^{(0)}, \mathbf{p}_{n,k}^{(+1)}\right) \quad (3)$$

$$\mathbf{c}_{n,k} = \left(c_{n,k}^{(-1)}, c_{n,k}^{(0)}, c_{n,k}^{(+1)}\right) \quad (4)$$

で与えられる．ここで，$k = 1 \ldots K$ は母音の開始や，アクセント句の開始など，時系列上に現れるイベントの種類を表す．本研究で用いるイベントをTable 1に示す．また $\mathbf{p}_{n,k}^{(u)}$，$c_{n,k}^{(u)}$ の $u \in \{-1, 0, +1\}$ は先行・当該・後続の音素やモーラといった言語単位のインデックスを表す．

$\mathbf{p}_{n,k}^{(u)}$ はフレームとイベントとの位置関係を表す位置コンテキストで，その次元数はイベントの種類によって異なる．例えば音素に関するイベントの場合，音素長が1になるように正規化した音素正規化位置と正規化を行わない時間からなる2次元ベクトル，アクセントに関するイベントの場合，イベントからのモーラ数を表すモーラ正規化位置，アクセント句正規化位置，および時間からなる3次元ベクトルを用いる．

$c_{n,k}^{(u)}$ はイベントの特徴量であり，音素の場合には音素弁別特性を表す $\{+1, -1\}$ のバイナリ特徴，モーラの場合には高低を表す $\{+1, -1\}$ のバイナリ特徴とする．一方でバイナリで表すことの困難な呼気段落の開始・終了や，アクセント核モーラの開始・終了などのイベント特徴量は1で固定とする．

Table 1 フレームコンテキストに用いるイベント

| | |
|---|---|
| 単位： | 音素 |
| イベント： | { 母音，高段，低段，前方，後舌，舌頂，破裂音，破擦音，継続音，有声，鼻音，半母音，無音 } 音素の { 開始，終了 } |
| 位置： | 音素正規化位置，時間 |
| 単位： | モーラ |
| イベント： | 高／低モーラの { 開始，終了 } |
| 位置： | { モーラ，アクセント句 } 正規化位置，時間 |
| 単位： | アクセント句 |
| イベント： | アクセント句の { 開始，終了 }<br>句頭音調モーラの { 開始，終了 }<br>アクセント核モーラの { 開始，終了 } |
| 位置： | { モーラ，アクセント句 } 正規化位置，時間 |
| 単位： | 呼気段落 |
| イベント： | 呼気段落の { 開始，終了 } |
| 位置： | { モーラ，アクセント句，呼気段落 } 正規化位置，時間 |
| 単位： | 文 |
| イベント： | 文の { 開始，終了 } |
| 位置： | { モーラ，アクセント句，呼気段落，文 } 正規化位置，時間 |

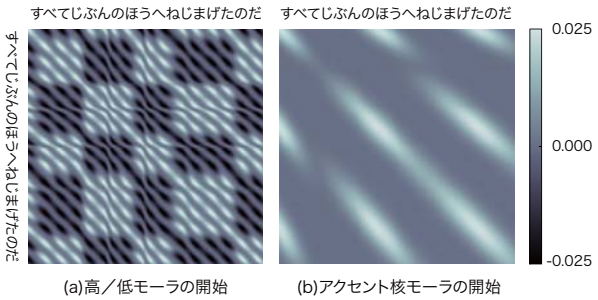


Fig. 1 イベントに関するカーネル行列の例

このコンテキストに対して，以下のカーネル関数

$$\kappa(\mathbf{x}_m, \mathbf{x}_n) = \sum_{k=1}^{K} \theta_{r,k}^2 \kappa_k(\mathbf{x}_{m,k}, \mathbf{x}_{n,k}) + \delta_{mn}\theta_{\text{floor}}^2 \quad (5)$$

$$\kappa_k(\mathbf{x}_{m,k}, \mathbf{x}_{n,k}) = \sum_{u=-1}^{+1}\sum_{v=-1}^{+1} w\left(\mathbf{p}_{m,k}^{(u)}\right) w\left(\mathbf{p}_{n,k}^{(v)}\right) \kappa_p\left(\mathbf{p}_{m,k}^{(u)}, \mathbf{p}_{m,k}^{(v)}\right) \kappa_c\left(c_{m,k}^{(u)}, c_{m,k}^{(v)}\right) \quad (6)$$

を定義する．$\theta_{a,k}^2$ は $k$ 番目のイベントに関するカーネル関数 $\kappa_k(\cdot)$ の重要度を表すハイパーパラメータで，$\delta_{mn}\theta_{\text{floor}}^2$ はカーネル関数の正定値性を保つための定数項である．式 (6) の右辺における総和は畳み込みカーネル [2] を表し，カーネル関数の言語単位間での連続性を保証する．また，$w(\cdot)$ はフレームの近くにあるイベントを重視するための重み関数であり，本研究ではガウス関数を用いる．$\kappa_p(\cdot)$，$\kappa_c(\cdot)$ はイベント

* A study on F0 contour generation based on Gaussian process regression. by KORIYAMA, Tomoki (Tokyo Institute of Technology), NOSE, Takashi (Tohoku University), and KOBAYASHI, Takao (Tokyo Institute of Technology)

Table 2　合成音声の有声／無声推定精度 (F 値)

| 手法＼話者 | MMY | FKS |
|---|---|---|
| HMM | 0.952 | 0.944 |
| GPR/GPC-PHONE | 0.963 | 0.956 |
| GPR/GPC-ALL | 0.964 | 0.957 |

の位置およびイベント特徴の類似度を表す部分カーネル関数であり，それぞれ Squared exponential (SE) カーネルと線形カーネルを用いる．

Fig. 1 にカーネル行列の例として，(a) 高／低モーラの開始および (b) アクセント核モーラの開始に関するカーネル行列を示す．図に示すように，アクセント句レベルのイベントであるアクセント核モーラの開始に関するカーネル関数は，モーラレベルのイベントである高/低モーラの開始に比べ，広いスケールでの相関関係を考慮している．このようにスケールの異なるカーネル関数を用いることで，F0 パタンの階層構造のモデル化が期待できる．

## 3　F0 パタンモデル

F0 パタンには値の存在する有声音のフレームと値のない無声音のフレームが存在するため，F0 パタンの生成には有声／無声を考慮したモデルが必要となる．本研究では，有声／無声推定モデルと F0 モデルの 2 種類のモデルを用いて F0 パタンのモデル化を行う．有声／無声推定モデルにはガウス過程分類 (GPC)[3] を用いる．ガウス過程分類とはロジスティック回帰によりガウス過程回帰と同様の枠組みでクラス分類を行う手法である．ガウス過程に従う潜在変数 $f(\mathbf{x})$ と有声／無声情報を表す観測変数 $y \in \{+1, -1\}$ の間に $p(y = +1) = \sigma(f(\mathbf{x}))$ の関係が成り立つとする．ただし，$\sigma(\cdot)$ はロジスティックシグモイド関数である．ガウス過程分類では予測分布を解析的に求められないため何らかの近似を行う必要があるが，本研究では計算コストの低いラプラス近似を用いる手法 [3] を利用する．F0 モデルには対数 F0 を特徴量とする GPR を使用し，学習データの有声フレームを用いて学習する．合成時には，まず有声／無声推定モデルを用いて合成する文の各フレームの有声／無声を決定する．そして，F0 モデルから有声フレーム区間における F0 の予測分布を求め，その平均系列を生成 F0 とする．

## 4　実験

### 4.1　実験条件

実験には ATR 日本語音声データベースセット B に含まれる男声話者，女声話者各 1 名 (MMY, FKS) 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を抽出し，それをもとに得られる 40 次元のメルケプストラム，対数 F0 および 5 次元の非周期性指標を音響特徴量とした．音響特徴量は次元ごとに平均 0，分散 1 となるように正規化を行った．

ガウス過程回帰・分類に用いる PIC 近似 [4] 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とした．カーネル関数のパラメータ最適化には EM アルゴリズムに基づく最適化法 [5] を使用した．このとき最適化には学習データの先頭の 50 文章を使用し，EM ステップの反復回数は 5 回とした．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用いられるコンテキスト決定木を使用するが，有声／無声推定モデルにはメルケプストラムの決定木を，F0 モデルには対数 F0 の決定木を使用した．音素継続長には原音声のものを用いた．

Table 3　合成音声の対数 F0 の RMS 誤差 [cent]

| 手法＼話者 | MMY | FKS |
|---|---|---|
| HMM | 220.0 | 192.6 |
| GPR/GPC-PHONE | 246.0 | 220.3 |
| GPR/GPC-ALL | 181.6 | 170.7 |

### 4.2　結果

有声／無声推定の精度を Table 2 に，合成音声の F0 歪を Table 3 に示す．表中の GPR/GPC-PHONE はフレームコンテキストに音素特徴だけを用いた場合，GPR/GPC-ALL はフレームコンテキストに Table 1 の全てのコンテキストを用いた場合にそれぞれ対応する．また，表中では従来手法として HMM 音声合成の結果を示している．有声／無声推定について，GPR/GPC-PHONE および GPR/GPC-ALL の精度は HMM より高く，GPC による有声／無声推定は効果的であることがわかる．また，F0 歪について，アクセント情報をフレームコンテキストに使用していない GPR/GPC-PHONE の F0 歪は HMM に比べ大きいが，GPR/GPC-ALL では F0 歪が大きく減少し HMM よりも小さくなっている．したがって，適切なコンテキストを導入することで，GPR に基づくフレームレベル音響モデルは F0 パタン生成に対しても有効であることがわかる．

## 5　おわりに

本稿ではガウス過程回帰・分類に基づく音声合成の枠組みにおいて，アクセントや呼気段落の情報をフレームコンテキストとして用いる方法を提案し，F0 パタンモデルの検討を行った．客観評価実験の結果から提案手法を用いることで，従来の HMM 音声合成に比べ有声／無声推定精度が向上し，合成音声の F0 歪が減少することを示した．今後は音素継続長のモデル化や主観評価実験による評価，適切なコンテキストの検討を行う予定である．

**謝辞**　本研究の一部は，日本学術振興会科学研究費補助金 (課題番号 24300071, 25540065, 25・8776) の助成を得た．

## 参考文献

[1] T. Koriyama et al.,, “Statistical Parametric Speech Synthesis Based on Gaussian Process Regression,” IEEE J-STSP, (8)2, pp.173–183.

[2] D. Haussler, “Convolution kernels on discrete structures,” Technical Report UCSC-CRL-99-10, Dept of Computer Science, UCSC, 1999.

[3] C.E. Rasmussen and C.K.I. Williams, *Gaussian processes for machine learning*, MIT press, 2006.

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

[5] 郡山他，“ガウス過程回帰に基づく音声合成におけるハイパーパラメータ最適化の検討，” 信学技報，vol.113, no.404，SP2013-99，pp.19–24，2014.