重回帰 HSMM に基づく音声合成における回帰行列の MAP 推定*

☆永田智洋, 森大毅 (宇都宮大), 能勢隆 (東工大)

1 はじめに

表情豊かな対話音声合成を実現するためには，言語情報だけではなくパラ言語情報を制御する必要がある。我々はこれまで，宇都宮大学パラ言語情報研究向け音声対話データベース (UUDB)[1] の感情の次元説に基づいて記述されたパラ言語情報についての抽象次元と，重回帰 HSMM に基づく音声合成手法 [2] を用いることで対話音声合成におけるパラ言語情報の制御が可能であることを報告した [3]。

しかし，重回帰 HSMM における回帰行列を学習する決定木のリーフノードに含まれる学習データが極端に少ないノードが存在するために，合成音声の自然性が著しく低くなるものが存在した。そこで，本研究では重回帰 HSMM における回帰行列の最大事後確率 (MAP) 推定を行い，ロバストな回帰行列の推定を行う。文献 [4] では，HSMM における平均パラメータの MAP 推定量から回帰行列を推定した。本研究では，HSMM における補助関数に回帰行列の事前分布を考慮することで，回帰行列の MAP 推定式を導出する。

2 重回帰 HSMM によるパラ言語情報制御

重回帰 HSMM では，HSMM の状態 i における出力確率分布の平均ベクトル $\boldsymbol{\mu}_i$ および状態継続長分布の平均 m_i が式 (1) で表せると仮定する。

$$\boldsymbol{\mu}_i = \boldsymbol{H}_{b_i}\boldsymbol{\xi}, \quad m_i = \boldsymbol{H}_{p_i}\boldsymbol{\xi} \tag{1}$$

$$\boldsymbol{\xi} = [1, v_1, v_2, \cdots, v_L]^\top = [1, \boldsymbol{v}^\top]^\top \tag{2}$$

$\boldsymbol{H}_{b_i} \in \boldsymbol{R}^{M\times(L+1)}$，$\boldsymbol{H}_{p_i} \in \boldsymbol{R}^{1\times(L+1)}$ はそれぞれ出力確率分布の平均ベクトル，状態継続長分布の平均に対する回帰行列である。ここで，$\boldsymbol{R}$ は実数の集合であり，L は重回帰モデルにおける独立変数で構成されるベクトル $\boldsymbol{v}$ の次元数，M は特徴ベクトルの次元数である。また，$\boldsymbol{\xi}$ は制御ベクトルである。

本研究では重回帰モデルの独立変数に，UUDB に記述されているパラ言語情報ラベルを用いることで合成音声のパラ言語情報の制御を行う。

$$\boldsymbol{\xi} = [1, v_{\mathrm{pl}}, v_{\mathrm{ar}}, \dots]^\top \tag{3}$$

ここで，$v_{\mathrm{pl}}, v_{\mathrm{ar}}, \cdots$ はそれぞれ「快-不快（pleasantness)」,「覚醒-睡眠（arousal)」，$\cdots$ の項目に対応する変数である。

3 回帰行列の MAP 推定

重回帰 HSMM の状態 i における出力確率分布の平均ベクトルに対する回帰行列 $\boldsymbol{H}_{b_i}$ の事前分布 $P(\boldsymbol{H}_{b_i})$ に，式 (4) に示す行列正規分布を仮定する。

$$P(\boldsymbol{H}_{b_i}) = (2\pi)^{\frac{M(L+1)}{2}} |\boldsymbol{\Omega}_{b_i}|^{-\frac{M}{2}} |\boldsymbol{\Phi}_{b_i}|^{-\frac{L+1}{2}} \cdot \exp\left\{-\frac{1}{2} tr\left(\boldsymbol{H}_{b_i} - \boldsymbol{W}_{b_i}\right)^\top \boldsymbol{\Omega}_{b_i}^{-1} \left(\boldsymbol{H}_{b_i} - \boldsymbol{W}_{b_i}\right) \boldsymbol{\Phi}_{b_i}^{-1}\right\} \tag{4}$$

ここで，$\boldsymbol{\Omega}_{b_i} \in \boldsymbol{R}^{M\times M}$ と $\boldsymbol{\Phi}_{b_i} \in \boldsymbol{R}^{(L+1)\times(L+1)}$，$\boldsymbol{W}_{b_i} \in \boldsymbol{R}^{M\times(L+1)}$ は MAP 推定における超パラメータである。

事前分布 $P(\boldsymbol{H}_{b_i})$ を HSMM における Q 関数に導入し，式 (5) に示す Q 関数を最大とする $\boldsymbol{H}_{b_i}$ を推定する。

$$Q(\lambda, \bar{b}_i) = \sum_{t=1}^{T}\sum_{d=1}^{t} \gamma_t^d(i) \sum_{s=t-d+1}^{t} \log \mathcal{N}(\boldsymbol{o}_s; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) + \log P(\boldsymbol{H}_{b_i}) \tag{5}$$

ここで，T は総フレーム数である。式 (5) を微分して 0 とおき，$\boldsymbol{\Omega}_{b_i} = \tau_{out}\boldsymbol{\Sigma}_i$ とすることで，$\boldsymbol{H}_{b_i}$ についての再推定式 (6) を得る。また，$\boldsymbol{H}_{p_i}$ に対しても同様の方法によって再推定式 (7) を得る。

$$\overline{\boldsymbol{H}}_{b_i} = \left\{\sum_{t=1}^{T}\sum_{d=1}^{t} \gamma_t^d(i) \sum_{s=t-d+1}^{t} \boldsymbol{o}_s\boldsymbol{\xi}^\top + \tau_{out}\boldsymbol{W}_{b_i}\boldsymbol{\Phi}_{b_i}\right\} \cdot \left\{\sum_{t=1}^{T}\sum_{d=1}^{t} \gamma_t^d(i)\cdot d\cdot \boldsymbol{\xi}\boldsymbol{\xi}^\top + \tau_{out}\boldsymbol{\Phi}_{b_i}\right\}^{-1} \tag{6}$$

$$\overline{\boldsymbol{H}}_{p_i} = \left\{\sum_{t=1}^{T}\sum_{d=1}^{t} \gamma_t^d(i)\cdot d\cdot \boldsymbol{\xi}^\top + \tau_{dur}\boldsymbol{W}_{p_i}\boldsymbol{\Phi}_{p_i}\right\} \cdot \left\{\sum_{t=1}^{T}\sum_{d=1}^{t} \gamma_t^d(i)\boldsymbol{\xi}\boldsymbol{\xi}^\top + \tau_{dur}\boldsymbol{\Phi}_{p_i}\right\}^{-1} \tag{7}$$

ここで，$\boldsymbol{o}_s$ は時刻 s における特徴ベクトルであり，T は特徴ベクトルの総フレーム数，$\gamma_t^d(i)$ は時刻 $t-d+1$ から t まで状態 i に滞在する確率であり，状態占有確率と呼ばれる。また，$\boldsymbol{W}_{p_i} \in \boldsymbol{R}^{1\times(L+1)}$ と $\boldsymbol{\Phi}_{p_i} \in \boldsymbol{R}^{(L+1)\times(L+1)}$ は $\boldsymbol{H}_{p_i}$ に対する MAP 推定時の超パラメータである。

4 客観評価実験

4.1 実験条件

学習には UUDB の対話セッション C002 から C007 に含まれる話者 FTS の 589 発話を用いた。550 発話の総時間は 17 分 25 秒である。スペクトルパラメータには，サンプリング周波数 16 kHz の音声信号から，分析周期 5 ms，分析窓長 25 ms のハミング窓を用いて求めた 0 次から 24 次のメルケプストラム係数を用いた。F0 パラメータは対数基本周波数とし，特徴ベクトルはこれらのパラメータにそれぞれの Δ，ΔΔ パラメータを加えた 78 次元のベクトルとした。回帰行列の再推定に必要となる初期回帰行列は文献 [5] による初期化手法を用いた。制御ベクトルは，感情状態を表す一般的な指標とされている「快-不快」,「覚醒-睡眠」を用いた式 (8) とし，各次元の値には UUDB

*Conversational speech synthesis with controllability of paralinguistic information using MRHSMM. by NAGATA, Tomohiro, MORI, Hiroki (Utsunomiya Univercity), NOSE Takashi (Tokyo Tech)

に記述されているラベラ 3 名による平均評価値を用いた。

$$\boldsymbol{\xi} = [1, v_{\mathrm{pl}}, v_{\mathrm{ar}}]^{\top} \tag{8}$$

回帰行列の学習法は，従来の最尤 (ML) 法と，式 (6) および式 (7) で示す MAP 推定法の 2 通りとした。重回帰 HSMM における回帰行列は，決定木に基づくクラスタリングを行った後に，決定木の各リーフノードに対して学習される。MAP 推定時の超パラメータ $\boldsymbol{W}_{b_i}$，$\boldsymbol{W}_{p_i}$ には，今回は回帰行列を学習するリーフノードにおいて ML 法で学習された回帰行列 $\boldsymbol{H}_{b_i}^{\mathrm{ML}(1)}$ と，その兄弟ノードにおいて ML 法で学習された回帰行列 $\boldsymbol{H}_{b_i}^{\mathrm{ML}(2)}$ を，それぞれのノードに含まれるコンテキストの種類による重み付き和で求めたものを用いた。これを式 (9)，式 (10) に示す。

$$\boldsymbol{W}_{b_i} = \frac{n_b^{(1)}}{n_b^{(1)} + n_b^{(2)}} \boldsymbol{H}_{b_i}^{\mathrm{ML}(1)} + \frac{n_b^{(2)}}{n_b^{(1)} + n_b^{(2)}} \boldsymbol{H}_{b_i}^{\mathrm{ML}(2)} \tag{9}$$

$$\boldsymbol{W}_{p_i} = \frac{n_p^{(1)}}{n_p^{(1)} + n_p^{(2)}} \boldsymbol{H}_{p_i}^{\mathrm{ML}(1)} + \frac{n_p^{(2)}}{n_p^{(1)} + n_p^{(2)}} \boldsymbol{H}_{p_i}^{\mathrm{ML}(2)} \tag{10}$$

ここで，$n_b^{(1)}$，$n_p^{(1)}$ は回帰行列を学習するノードに含まれるコンテキストの種類数であり，$n_b^{(2)}$，$n_p^{(2)}$ は兄弟ノードに含まれるコンテキストの種類数である。また，$\boldsymbol{\Phi}_{b_i}$，$\boldsymbol{\Phi}_{p_i}$ はそれぞれ単位行列とした。

テストセットは UUDB の対話セッション C001 の話者 FTS の 95 発話とした。

4.2 客観評価

ML 法によって合成された音声と MAP 法によって合成された音声を比較した結果，不自然な音響特徴量が出力された結果が減少していた。例として,「快-不快」に 4,「覚醒-睡眠」に 2 を与えて合成された発話「で，ヘックションっつって，あ急に寒くなったっつったって」の対数基本周波数軌跡を図 1 に示す。MAP による結果では，極端な対数基本周波数が出力されていないことがわかる。

次に，合成時に与えたパラ言語情報と合成された音声の音響特徴量を評価した。合成時に付与した「快-不快」,「覚醒-睡眠」の値と合成された音声の音響特徴量の相関係数を示す。表 1 は ML 法による結果であり，表 2 は MAP 法による結果である。ここで,「快-不快」の相関係数は，合成時に付与する「覚醒-睡眠」の値を 4 として「快-不快」の値を変化させたときの相関係数であり,「覚醒-睡眠」の相関係数は，合成時に付与する「快-不快」の値を 4 として「覚醒-睡眠」の値を変化させたときの相関係数である。また，音響特徴量は，発話における基本周波数最大値 (F0 最大値) および基本周波数平均値 (F0 平均値) とした。

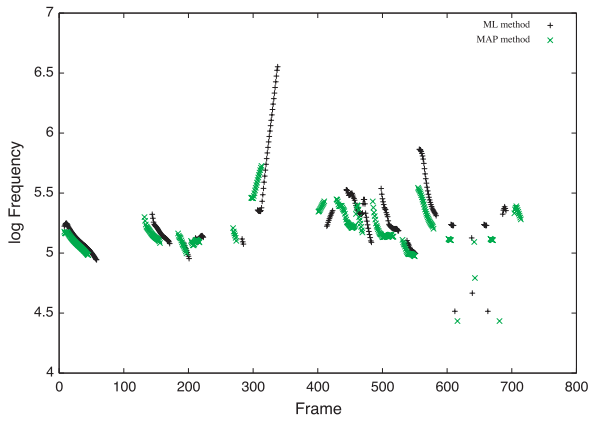

Fig. 1　発話の対数基本周波数軌跡

Table 1　ML 法におけるパラ言語情報の値と合成音声の音響特徴量との相関係数

	快-不快	覚醒-睡眠
F0 最大値	0.02	0.41
F0 平均値	0.00	0.89

Table 2　MAP 法におけるパラ言語情報の値と合成音声の音響特徴量との相関係数

	快-不快	覚醒-睡眠
F0 最大値	0.02	0.57
F0 平均値	0.00	0.91

表より，ML 法と MAP 法のどちらの場合においても「覚醒-睡眠」と音響特徴量との間には高い相関があることが確認できた。しかし,「快-不快」と音響特徴量の間には相関が確認されなかった。この原因を調べるために，発話ごとに「快-不快」と音響特徴量との相関係数を求めた。その結果，F0 最大値の場合には正の相関を持つ発話が 52 発話あり，負の相関を持つ発話が 43 発話，F0 平均値の場合には正の相関を持つ発話が 47 発話あり，負の相関を持つ発話が 48 発話あった。また，どの発話についても強い相関が得られていた。このことから,「快-不快」を変化させたことによる音響特徴量の変化は，発話内容に依存し,「覚醒-睡眠」を変化させたことによる音響特徴量の変化は発話内容に依存しないことが確認できた。

5　おわりに

本研究では，重回帰 HSMM に基づく音声合成における回帰行列の MAP 推定を行った。合成時に付与したパラ言語情報と合成音声の音響特徴量との関係を評価し，各音響特徴量との依存性を調べた。

今後は，合成された音声を用いた主観評価実験を行い，合成音声の自然性を評価する。また，意図したパラ言語情報が表現可能であるかを示すために，パラ言語情報の表出実験を行う。

参考文献

[1] Mori *et al.*, Speech Communication, 53, 36-50, 2011.
[2] Nose *et al.*, IEICE Trans. Inf. & Syst., E90-D(9), 1406-1413, 2007.
[3] 永田 他, 音講論 (春), 435–436, 2012.
[4] Nose *et al.*, IEICE Trans. Inf. & Syst., E92-D(3), 489-497, 2009
[5] 能勢, 小林, 音講論 (秋), 329–330, 2011.