Vol.2010-SLP-84 No.29
2010/12/21

HMM 音声合成における系列内変動モデリング手法の評価

是竹有里[†1]　戸田智基[†1]　木佐木雄介[†1]
猿渡洋[†1]　鹿野清宏[†1]

隠れマルコフモデル（hidden Markov model: HMM）に基づく音声合成は，様々な話者，発話様式，言語へ容易に適用できるなど，柔軟性に優れた音声合成処理を実現できる．一方で，HMM から生成される音声パラメータは，汎化処理の影響により過剰に平滑化されたものとなるため，一般的に合成音声の自然性は劣化する．この問題点を緩和する手法として，系列内変動（global variance: GV）モデリングが提案されている．生成パラメータの GV は一般的に大きく減少する傾向があるため，適切な GV を持つパラメータを生成することで，過剰な平滑化を効果的に抑制できる．本報告では，代表的な手法として，パラメータ生成時に GV を考慮する手法と，HMM の学習時から GV を考慮する手法に着目し，これら二つの手法を様々な観点から比較することで，GV モデリングの効果について考察する．

An evaluation of modeling methods of global variance in HMM-based speech synthesis

Yuri Koretake,[†1] Tomoki Toda ,[†1] Yusuke Kisaki,[†1]
Hiroshi Saruwatari[†1] and Kiyohiro Shikano[†1]

Speech synthesis based on hidden Markov model (HMM) is a technique capable of flexibly developing various Text-to-Speech voices, such as various speakers, various speaking styles, various languages, and so on. On the other hand, naturalness of synthetic speech is usually degraded by the use of the over-smoothed speech parameters generated from HMMs. To address this issue, some modeling methods of global variance (GV) have been proposed. Since the GV is inversely correlated with the over-smoothing effect, it is effectively alleviated by the use of a metric on the GV of the generated parameters. In this report, we focus on two typical GV-modeling methods: 1) a parameter generation method considering the GV and 2) a GV-constrained HMM training method. The effects of the GV modeling are investigated by a comparison of these two methods from various perspectives.

1. はじめに

任意の文字情報から音声を生成する技術として，テキスト音声合成 (Text-To-Speech: TTS) がある．1990 年頃に考案されたコーパスベース方式[1] は，予め収録された音声データを利用して，機械学習等の数理的なアルゴリズムにより，TTS の個々の処理を半自働的に構築する．汎用性に優れた技術であり，近年の主流方式として，広く研究されている．

本稿では，コーパスベース方式の一つとして，隠れマルコフモデル (hidden Markov model: HMM) に基づくテキスト音声合成に着目する．HMM 音声合成は統計的パラメトリック音声合成方式であり，音声パラメータの統計量を用いた音声合成処理を行う[2]．まず，学習処理として，音声コーパスを用いて，音声パラメータ系列をモデル化する HMM を学習する．そして，合成処理では，入力テキストに対応する HMM から，尤度最大化基準によって音声パラメータを生成する．動的特徴量を考慮することで，適切に遷移するパラメータ系列を生成することができ，不連続感の少ない安定した品質の合成音声が得られるという特徴がある[3]．様々な発話様式，言語に対して容易に適用可能であり，モデル適応処理により少量のデータで様々な音声を容易に合成することも可能である．一方で，汎化処理の影響により，HMM から生成されるパラメータは過剰に平滑化されたものとなり，合成音声の自然性が大きく劣化するという問題点がある．

この問題に対して，HMM の学習処理を改善する手法や，合成時に素片選択処理を併用するハイブリッド方式[4] が提案されている．学習処理を改善する代表的な手法として，最小生成誤差学習法[5] やトラジェクトリ学習法[6] がある．通常の HMM 音声合成の枠組みでは，学習時には静的・動的結合特徴量系列に対して最適化を行うのに対し，合成時には静的特徴量系列に対して最適化を行うため，学習時と合成時の最適化基準が一致しない．これに対して，最小生成誤差学習法では，学習パラメータと生成パラメータの誤差が最小になるように HMM パラメータを最適化する．また，トラジェクトリ学習法では，静的・動的特徴量間の関係を HMM に導入することで導出されるトラジェクトリ HMM を用いて，静的特徴量系列に対する最適化を行う．どちらの手法も，通常の枠組みと比較し，合成音声の自然性を改善できるものの，その改善効果は十分ではない．一方で，ハイブリッド方式では，HMM から生成されるパラメータを用いるのではなく，音声コーパス中の波形素片を用いて，音声波形を合成する．合成音声の自然性を大きく改善することができるが，十分な品質を確保するための音声データ量は増加する．また，モデル適応処理などを直接使用することが困難となる．

効果的かつ効率的に過剰な平滑化を緩和できる手法として，系列内変動 (global variance: GV) モデリングが提案されている[7]．GV とは，一発話単位など，パラメータ時系列全体における静的特徴量の分散のことである．過剰な平滑化の影響により，HMM から生成される音声パラメータの GV は大きく減少する傾向がある．GV を考慮したパラメータ生成法[8] では，学習データの各発話から計算される GV を用いて，HMM とは別に，GV の確率密

†1 奈良先端科学技術大学院大学
Nara Institute of Science and Technology

Vol.2010-SLP-84 No.29
2010/12/21

度をモデル化する分布パラメータを学習する．パラメータ生成時に，HMM の尤度だけでなく，GV の分布の尤度も考慮することで，適切な GV を持つパラメータ系列を生成する．これにより，合成音声の品質は大幅に改善される．さらに，学習時にも GV を考慮することで同一の基準による学習・合成処理を実現する手法として，GV を考慮したトラジェクトリ学習法が提案されている[9]．この手法では，コンテキスト依存分布で GV の確率密度分布をモデル化できるなど，GV を考慮したパラメータ生成法と比較して，いくつかの利点が存在する．

本稿では，合成音声の自然性を大きく改善する効果のある GV のモデリング手法に関して，詳細な検討を行う．GV に基づくパラメータ生成法と GV を考慮したモデル学習法に対して，合成音声の自然性や生成パラメータの GV の統計的性質といった観点から比較を行い，各手法の特徴を明らかにする．また，GV を用いる際に適したスペクトル特徴量を明らかにするため，極零型であるメルケプストラム係数と全極型であるメル LSP 係数を用いた際の GV の効果の違いについても調査する．

以下，2 節では HMM 音声合成について述べ，3 節では GV を考慮したパラメータ生成，4 節では GV を考慮したモデル学習について述べる．そして，5 節では実験的評価を述べ，6 節で本稿のまとめを述べる．

2. HMM 音声合成

2.1 モデル学習

HMM 音声合成では，スペクトル特徴量や音源特徴量といった音声パラメータの時系列を HMM でモデル化する．フレーム t における観測ベクトル $\boldsymbol{o}_t = \left[\boldsymbol{c}_t^\mathsf{T}, \Delta^{(1)}\boldsymbol{c}_t^\mathsf{T}, \Delta^{(2)}\boldsymbol{c}_t^\mathsf{T}\right]^\mathsf{T}$ は，D 次元の静的特徴量ベクトル $\boldsymbol{c}_t = [c_t(1), \cdots, c_t(D)]^\mathsf{T}$ と，その 1 次と 2 次の動的特徴量ベクトル $\Delta^{(1)}\boldsymbol{c}_t, \Delta^{(2)}\boldsymbol{c}_t$ の結合ベクトルで構成される．$\boldsymbol{o}_t$ と $\boldsymbol{c}_t$ のベクトル系列をそれぞれ $\boldsymbol{o} = \left[\boldsymbol{o}_1^\mathsf{T}, \cdots, \boldsymbol{o}_t^\mathsf{T}, \cdots, \boldsymbol{o}_T^\mathsf{T}\right]^\mathsf{T}, \boldsymbol{c} = \left[\boldsymbol{c}_1^\mathsf{T}, \cdots, \boldsymbol{c}_t^\mathsf{T}, \cdots, \boldsymbol{c}_T^\mathsf{T}\right]^\mathsf{T}$ として表し，HMM の状態系列を $\boldsymbol{q} = [q_1, \cdots, q_t, \cdots, q_T]$ とすると，観測ベクトル系列は次式にてモデル化される．

$$P(\boldsymbol{o}|\boldsymbol{\lambda}) = \sum_{all\boldsymbol{q}} P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\lambda}) P(\boldsymbol{q}|\boldsymbol{\lambda}) \tag{1}$$

ここで，

$$P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_{\boldsymbol{q}}, \boldsymbol{U}_{\boldsymbol{q}}) = \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_{qt}, \boldsymbol{U}_{qt}) \tag{2}$$

$$\boldsymbol{\mu}_{\boldsymbol{q}} = [\boldsymbol{\mu}_{q1}^\mathsf{T}, \cdots, \boldsymbol{\mu}_{qt}^\mathsf{T}, \cdots \boldsymbol{\mu}_{qT}^\mathsf{T}]^\mathsf{T} \tag{3}$$

$$\boldsymbol{U}_{\boldsymbol{q}} = \mathrm{diag}[\boldsymbol{U}_{q1}, \cdots, \boldsymbol{U}_{qt}, \cdots \boldsymbol{U}_{qT}] \tag{4}$$

であり，$\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{U})$ は平均ベクトル $\boldsymbol{\mu}$，共分散行列 $\boldsymbol{U}$ の正規分布を表す．また，状態継続長を制御するために，$P(\boldsymbol{q}|\boldsymbol{\lambda})$ は状態遷移確率と状態継続長分布を用いてモデル化する[10]．

HMM パラメータセット $\boldsymbol{\lambda}$ は，以下に示す通り，最尤基準により最適化する．

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \sum_{all\boldsymbol{q}} P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\lambda}) P(\boldsymbol{q}|\boldsymbol{\lambda}) \tag{5}$$

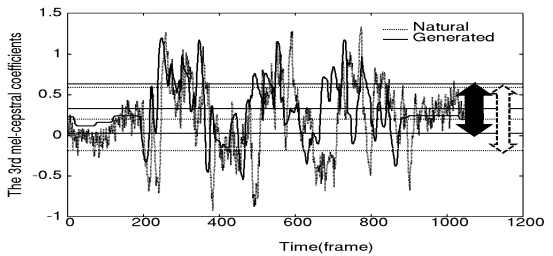

図 1 自然音声と生成音声のメルケプストラム系列．右端の矢印は，各系列の GV の大きさを表す．
Fig. 1 Natural and generated mel-cepstrum sequences. Square root of GV of each sequence is shown by bidirectional arrows.

2.2 パラメータ生成

パラメータ生成では，与えられた言語情報に基づき HMM を構成し，状態継続長分布の尤度に基づき状態系列 $\boldsymbol{q} = [q_1, \cdots, q_t, \cdots, q_T]$ を決定する．その後，音声パラメータの静的特徴量ベクトル系列を次式にて生成する．

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\lambda}) = \bar{\boldsymbol{c}}_{\boldsymbol{q}} \qquad \text{subject to } \boldsymbol{o} = \boldsymbol{W}\boldsymbol{c} \tag{6}$$

ここで，$\boldsymbol{W}$ は静的特徴量系列から静的・動的特徴量系列へ変換するための行列を表す．最尤系列 $\bar{\boldsymbol{c}}_{\boldsymbol{q}}$ は以下に示すとおり，解析的に求められる．

$$\bar{\boldsymbol{c}}_{\boldsymbol{q}} = \boldsymbol{P}_{\boldsymbol{q}}\boldsymbol{r}_{\boldsymbol{q}} \tag{7}$$

$$\boldsymbol{P}_{\boldsymbol{q}}^{-1} = \boldsymbol{W}^\mathsf{T}\boldsymbol{U}_{\boldsymbol{q}}^{-1}\boldsymbol{W} \tag{8}$$

$$\boldsymbol{r}_{\boldsymbol{q}} = \boldsymbol{W}^\mathsf{T}\boldsymbol{U}_{\boldsymbol{q}}^{-1}\boldsymbol{\mu}_{\boldsymbol{q}} \tag{9}$$

3. GV を考慮したパラメータ生成

一般的に，HMM から生成される最尤系列 $\bar{\boldsymbol{c}}_{\boldsymbol{q}}$ は汎化処理の影響により，過剰に平滑化されたものとなる．そのような音声パラメータの使用は，合成音声の自然性を大幅に劣化させる．この問題を緩和する方法として，GV を考慮したパラメータ生成法が提案されている．

3.1 GV

静的特徴量ベクトルの GV は次式にて定義される．

$$\boldsymbol{v}(\boldsymbol{c}) = [v(1), v(2), ..., v(D)]^\mathsf{T} \tag{10}$$

$$v(d) = \frac{1}{T}\sum_{t=1}^{T} \left(c_t(d) - \bar{c}(d)\right)^2 \tag{11}$$

$$\bar{c}(d) = \frac{1}{T}\sum_{\tau=1}^{T} c_\tau(d) \tag{12}$$

本稿では，1 発話ごとに GV を計算する．

図 1 に，自然音声から抽出されたメルケプストラム係数の時系列と，HMM から最尤基

Vol.2010-SLP-84 No.29
2010/12/21

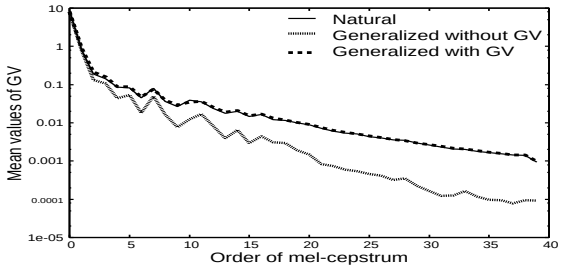

図 2 GV の平均値の比較
Fig. 2 Comparison of mean values of GV between some parameter sequences

準で生成されたメルケプストラム係数の時系列を示す．生成されたメルケプストラム係数は自然音声のメルケプストラム係数に比べ，GV が大きく減少する傾向が見られる．

3.2 パラメータ生成法

適切な GV を持つ音声パラメータを生成するために，$\boldsymbol{o}$ に対する尤度のみでなく，$\boldsymbol{v}(\boldsymbol{c})$ に対する尤度も考慮して，以下のように静的特徴量系列を決定する．

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} P(\boldsymbol{o}|\boldsymbol{q},\boldsymbol{\lambda}) P(\boldsymbol{v}(\boldsymbol{c})|\boldsymbol{\lambda}_v)^{\omega T} \qquad \text{subject to } \boldsymbol{o} = \boldsymbol{W}\boldsymbol{c} \tag{13}$$

ここで，定数 ω は二つの尤度間の重みを調節するパラメータであり，手動で設定する．GV の確率密度を平均ベクトル $\boldsymbol{\mu}_v$，共分散行列 $\boldsymbol{\Sigma}_v$ の正規分布 $\boldsymbol{\lambda}_v$ によりモデル化する．なお，式 (13) を解析的に解くことは困難であるため，勾配法による繰り返し処理が必要となる．

3.3 効　　果

図 2 に自然音声から抽出されたメルケプストラム，GV を考慮せずに生成されたメルケプストラム (式 (7)) ，および，GV を考慮して生成されたメルケプストラム (式 (13)) の GV の平均値を示す．GV を考慮せずに生成されたパラメータの GV は，自然音声のものと比較し，減少していることが分かる．また，その差は，高次の係数になるにつれて顕著である．それに比べ，GV を考慮したパラメータ生成では，GV の尤度が制約条件として機能するため，生成されるパラメータの GV は，自然音声のものとほぼ等しくなることが分かる．

なお，GV を考慮したパラメータ生成により得られる合成音声の自然性改善効果は，音声パラメータに依存する傾向がある．例えば，対数 F_0 に対する改善効果は非常に小さいが，スペクトル特徴量に対する改善効果は大きい傾向がある．[8)9)] また，スペクトル特徴量においても，ケプストラムなどの極零型パラメータと，LSP などの全極型パラメータで，改善効果が異なる傾向がある．[11)]

4. GV を考慮したモデル学習

GV を考慮したパラメータ生成法により，合成音声の自然性は大幅に改善されるが，生成パラメータを解析的に求めることができず，反復処理が必要となる．また，HMM と GV の確率密度分布を独立に学習し，合成時には式 (13) に示すようにそれらの結合確率の最大化を行うため，学習時と合成時の最適化基準が一致しないという問題がある．さらに，GV の確率密度分布が極めて単純であるため，コンテキストの変化に対応できない．そのため，例えば，非常に短い文を合成する際など，不自然な音声パラメータが生成されることがある．図 3 および図 4 に，GV 計算に用いるフレーム数を変化させた際に生じるメルケプストラム係数の GV の分布変化と，対数 F_0 の GV の分布変化を示す．フレーム数によって，GV の値が変化する傾向が見られる．GV に対して，コンテキスト依存確率密度分布を用いることも考えられるが，発話単位で GV を計算する際には，学習時に得られる GV のサンプル数は高々発話数までであるため，頑健に学習できる分布数は限られる．

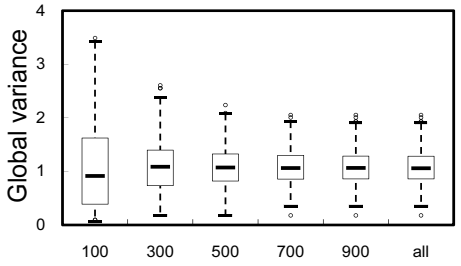

図 3 フレーム数の違いによる一次のメルケプストラム係数の GV の変化
Fig. 3 GV mean values of the 1st mel-cepstral coefficient as a funcion of the number of frames.

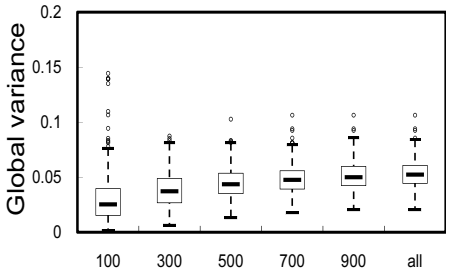

図 4 フレーム数の違いによる対数 F_0 の GV の変化
Fig. 4 GV mean values of log-scaled F_0 as a funcion of the number of frames.

これらの問題に対して，パラメータ生成時のみでなく，学習時にも GV を考慮する方法として，GV 制約付きトラジェクトリ学習法が提案されている．

4.1 学　習　法

GV 制約付きトラジェクトリ学習法では，学習時と合成時に同一の最適化基準が用いられる．通常の HMM に対して，最適化基準の同一化を可能とする手法がトラジェクトリ学習[6)] であり，その枠組みにおいて，さらに GV を考慮することを可能にしたものが，GV 制約付きトラジェクトリ学習である．ここでは，まずトラジェクトリ学習について説明した後に，GV 制約付きトラジェクトリ学習について述べる．

式 (6) の条件部で示される合成部に考慮する静的特徴量と動的特徴量の明示的な関係を，HMM に対して導入することで，トラジェクトリ HMM が導出される．トラジェクトリ学習法は，トラジェクトリ HMM の尤度が最大となるように，HMM パラメータを最適化する手法である．状態系列 $\boldsymbol{q}$ が与えられた際の，トラジェクトリ HMM による静的特徴量系

Vol.2010-SLP-84 No.29
2010/12/21

列 $\boldsymbol{c}$ の確率密度分布は，以下のように表される．

$$P(\boldsymbol{c}|\boldsymbol{q},\boldsymbol{\lambda}) = \frac{1}{\int P(\boldsymbol{W}\boldsymbol{c}|\boldsymbol{q},\boldsymbol{\lambda})d\boldsymbol{c}} P(\boldsymbol{o}|\boldsymbol{q},\boldsymbol{\lambda}) \tag{14}$$

$$= \frac{1}{Z_{\boldsymbol{q}}} P(\boldsymbol{o}|\boldsymbol{q},\boldsymbol{\lambda}) \tag{15}$$

$$= \mathcal{N}(\boldsymbol{c};\bar{\boldsymbol{c}}_{\boldsymbol{q}},\boldsymbol{P}_{\boldsymbol{q}}) \tag{16}$$

ここで，正規化定数 $Z_{\boldsymbol{q}}$ は次式で表される．

$$Z_{\boldsymbol{q}} = \frac{\sqrt{(2\pi)^{DT}|\boldsymbol{P}_{\boldsymbol{q}}|}}{\sqrt{(2\pi)^{3DT}|\boldsymbol{U}_{\boldsymbol{q}}|}} \exp\left(-\frac{1}{2}(\boldsymbol{\mu}_{\boldsymbol{q}}^T\boldsymbol{U}_{\boldsymbol{q}}^{-1}\boldsymbol{\mu}_{\boldsymbol{q}} - \boldsymbol{r}_{\boldsymbol{q}}^T\boldsymbol{P}_{\boldsymbol{q}}\boldsymbol{r}_{\boldsymbol{q}})\right) \tag{17}$$

トラジェクトリ学習法では，以下に示す通り，HMM パラメータセット $\boldsymbol{\lambda}$ はトラジェクトリ HMM の尤度が最大になるように最適化される．

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} P(\boldsymbol{c}|\boldsymbol{q},\boldsymbol{\lambda}) \tag{18}$$

GV 制約付きトラジェクトリ学習[9] では，状態系列 $\boldsymbol{q}$ が与えられた際に，HMM パラメータセット $\boldsymbol{\lambda}$ は以下のように最適化される．

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} P(\boldsymbol{c}|\boldsymbol{q},\boldsymbol{\lambda}) P(\boldsymbol{v}(\boldsymbol{c})|\boldsymbol{q},\boldsymbol{\lambda},\boldsymbol{\lambda}_v)^{\omega T} \tag{19}$$

ここで，学習基準は，トラジェクトリ HMM の尤度と次式で表される GV の確率密度分布の尤度の積で与えられる．

$$P(\boldsymbol{v}(\boldsymbol{c})|\boldsymbol{q},\boldsymbol{\lambda},\boldsymbol{\lambda}_v) = \mathcal{N}(\boldsymbol{v}(\boldsymbol{c});\boldsymbol{v}(\bar{\boldsymbol{c}}_{\boldsymbol{q}}),\boldsymbol{\Sigma}_{\boldsymbol{v}}) \tag{20}$$

GV の確率密度分布の平均ベクトルは，トラジェクトリ HMM の平均ベクトル (式 (6) で与えられる最尤系列と等価) の GV として定義される．そのため，状態系列の変化，すなわち，コンテキストの変化に応じて，GV の確率密度分布の平均ベクトルは変化する．パラメータ生成の際には，次式に示す通り，学習時と同一の評価基準が用いられる．

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{\lambda}} P(\boldsymbol{c}|\boldsymbol{q},\boldsymbol{\lambda}) P(\boldsymbol{v}(\boldsymbol{c})|\boldsymbol{q},\boldsymbol{\lambda},\boldsymbol{\lambda}_v)^{\omega T} \tag{21}$$

$$= \arg\max_{\boldsymbol{c}} \mathcal{N}(\boldsymbol{c};\bar{\boldsymbol{c}}_{\boldsymbol{q}},\boldsymbol{P}_{\boldsymbol{q}}) \mathcal{N}(\boldsymbol{v}(\boldsymbol{c});\boldsymbol{v}(\bar{\boldsymbol{c}}_{\boldsymbol{q}}),\boldsymbol{\Sigma}_v)^{\omega T} \tag{22}$$

$$= \bar{\boldsymbol{c}}_{\boldsymbol{q}} \tag{23}$$

この時の最尤系列は解析的に求めることができる．

4.2 効　　果

図 5 に発話単位で計算されたメルケプストラムの GV の分布を示し，図 6 に対数 F_0 の GV の分布を示す．トラジェクトリ学習および GV を考慮しないパラメータ生成を行う際には，自然音声と比べ，GV の減少が見られる．GV を考慮したパラメータ生成により，GV の減少は抑えられるが，生成パラメータの GV は，コンテキストに関わらず，ほぼ一定の値となる．一方，GV 制約付きトラジェクトリ学習法の場合は，合成音声の GV は，自然音声の GV と比較的類似した分布形状を持つことが分かる．

5. 実験的評価

5.1 実 験 条 件

話者として，日本人女性話者 1 名による音声データを用いる．学習データとして ATR 音素バランス文 503 文のうち A から I セット 450 文を用い，評価データとして J セット 53 文を用いる．サンプリング周波数は 16 kHz，フレームシフトは 5 ms とする．音声分析法として，STRAIGHT[12] を用いる．スペクトル特徴量は，メルケプストラムまたはメル LSP 及び対数ゲインを用いる．分析次数は 39，周波数軸伸縮パラメータは 0.42 とする．音源特徴量として，対数 F_0 と 5 帯域で平均化された非周期成分を用いる．

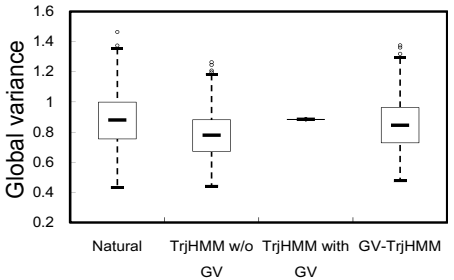

図 5 1 次のメルケプストラム係数の GV の分布
Fig. 5 Distribution of GV of the 1st mel-cepstral coefficient

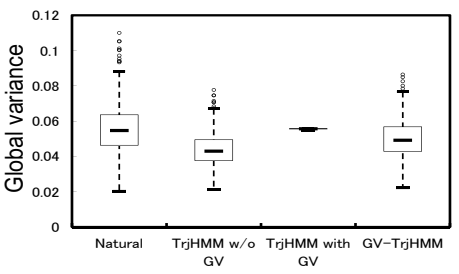

図 6 対数 F_0 の GV の分布
Fig. 6 Distribution of GV of log-scaled F_0

スペクトル，非周期成分，対数 F_0 は，それぞれ別ストリームとして，5 状態 left-to-right 型のコンテキスト依存音素 HMM でモデル化する．対数 F_0 に関しては，多空間上の確率分布に基づく HMM[13] を用いる．最小記述長（minimum description length: MDL）基準によるコンテキストクラスタリング[14] を用いて，コンテキスト依存音素 HMM を作成する．最尤基準により，状態継続長分布付き HMM を学習した後に，HMM による Viterbi アライメントにより状態系列を決定し，トラジェクトリ学習，さらには，GV 制約付きトラジェクトリ学習を行う．GV 制約付きトラジェクトリ学習における GV 重みは 0.1 とする．合成時には，HMM から生成されるパラメータに基づき，STRAIGHT 混合励振源[15] および MLSA フィルタ[16] を用いて，音声波形を合成する．

5.2 各手法の比較

5.2.1 実 験 条 件

GV を考慮したパラメータ生成法と GV 制約付きトラジェクトリ学習法を比較するために，合成音声の自然性に関するオピニオン評価を行う（実験 1）．各スペクトルパラメータにおいて GV を考慮する効果も明らかにするため，メルケプストラムおよびメル LSP に対して，1) 標準的な学習法による合成音声 [HMM w/o GV]，2) 標準的な学習法および GV を考慮したパラメータ生成による合成音声 [HMM with GV]，3) トラジェクトリ学習法による合成音声 [TrjHMM w/o GV]，4) トラジェクトリ学習法および GV を考慮したパラメータ生成による合成音声 [TrjHMM with GV]，5) GV 制約付きトラジェクトリ学習法による合成音声 [GV-TrjHMM] を作成する．評価には，これらの合成音声と分析合成音声 [Ana-syn] から成る計 12 種類の音声サンプルを用いる．被験者は 10 名であり，被験者ごとに各種音

Vol.2010-SLP-84 No.29
2010/12/21

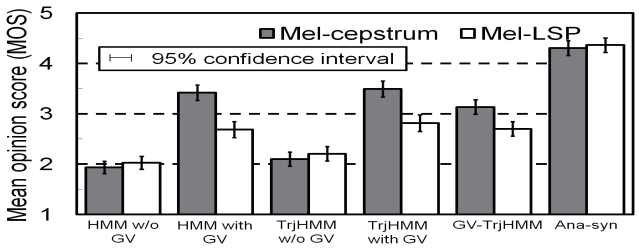

図 7 自然性に関する主観評価の結果
Fig. 7 Result of opinion test on naturalness

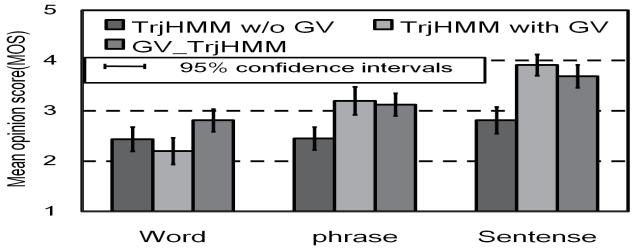

図 8 文長を変化させた際の自然性に関する主観評価結果
Fig. 8 Result of opinion test on naturalness when varying sentence length.

声に対して 53 文の中からランダムに選ばれた 15 文，計 180 文を 5 段階オピニオンスコアで評価する．

5.2.2 自然性に関する主観評価の結果

実験 1 の結果を図 7 に示す．スペクトル特徴量や学習法に関わらず，GV を考慮することで自然性は大きく改善する．メル LSP よりもメルケプストラムの方が，GV を考慮することによる改善効果が大きく．スペクトル特徴量の違いによって GV の効果が変わることが分かる．また，GV 制約付きトラジェクトリ学習は，GV を考慮したパラメータ生成とほぼ同程度の音質改善効果が得られるが，メルケプストラムを用いる際には若干劣化する傾向が見られる．

5.3 文長が与える影響

5.3.1 実 験 条 件

GV を考慮したパラメータ生成法と GV 制約付きトラジェクトリ学習法の効果をより詳細に比較するために，様々な長さの文に対する合成音声の自然性を評価する（実験 2）．3 種類の異なる長さの文（例:奈良 [Word]，奈良は最高 [Phrase]，奈良には世界遺産が沢山あります [Sentence]）をそれぞれ 10 セット用意し，計 30 文に対して音声合成を行う．スペクトル特徴量にはメルケプストラムを用い，1) トラジェクトリ学習による合成音声 [TrjHMM w/o GV]，2) トラジェクトリ学習および GV を考慮したパラメータ生成による合成音声 [TrjHMM with GV]，3) GV 制約付きトラジェクトリ学習による合成音声 [GV-TrjHMM] を作成する．被験者は 8 名であり，被験者ごとに各種合成音声に対して 10 文，計 90 文を 5 段階オピニオンスコアで評価する．

5.3.2 文長を変化させた際の自然性に関する主観評価結果

実験 2 の結果を図 8 に示す．短い文のコンテキストは学習データに含まれていないため，全体的に文長が短くなるにつれ，自然性は劣化する傾向が見られる．GV を考慮したパラメータ生成においては，その劣化が大きい傾向がある．これは，コンテキスト非依存モデルで GV の確率密度分布をモデル化しているため，文長が短くなるにつれ，実際の分布とのずれが大きくなるためである．一方で，GV 制約付きトラジェクトリ学習においては，コンテキストに応じた GV の確率密度分布を用いることができるため，短い文を合成する際においても，GV を考慮する効果が見られる．

各手法において，文長が GV の確率密度分布に与える影響を調査するために，音声開始から様々な長さの音声区間を抜き出すことで，模擬的に文長の異なる音声に対する GV を計算する．図 9 から図 14 に，1) 自然音声 [Natural]，2) トラジェクトリ学習による合成音声 [TrjHMM w/o GV]，3) トラジェクトリ学習および GV を考慮したパラメータ生成による合成音声 [TrjHMM with GV]，4) GV 制約付きトラジェクトリ学習による合成音声 [GV-TrjHMM] におけるメルケプストラムの GV の分布を示す．ここでは，GV 計算時には，無音フレームを除いている．GV を考慮したパラメータ生成では，コンテキスト非依存の確率分布を使用するため，GV の分散は小さくなる．また，図中の中央値に注目すると，フレーム数が減少するにつれ，自然音声と比較してより大きくなる傾向がある．一方で，GV 制約付きトラジェクトリ学習法では，より自然音声に近い GV の分布が得られることが分かる．

以上の結果から，GV 制約付きトラジェクトリ学習では，様々な文長に対しても，安定した自然性改善効果が得られることが分かる．ただし，図 7 に見られるように，通常の長さの文を合成した際には，GV を考慮したパラメータ生成法と比べて，若干自然性改善効果が低くなる原因については，今後さらに調査する必要がある．

6. ま と め

本報告では，HMM 音声合成において，合成音声の自然性改善に有効な特徴量として系列内変動（global variance: GV）に注目し，GV を考慮したパラメータ生成法と GV 制約付きトラジェクトリ学習法の比較を通して，GV のモデリング手法に関する検討を行った．GV の確率密度分布は，合成する文の長さが短くなると大きく変化する傾向があることが分かった．合成音声の自然性に関する主観評価実験の結果，GV を考慮したパラメータ生成法では，極端に短い文長を合成する際には，自然性改善効果が得られないことが分かった．一方で，GV 制約付きトラジェクトリ学習法では，コンテキストに応じて GV の確率密度分布を変化させることができるため，様々な文長を合成する際においても，安定した自然性改善効果が得られることが分かった．また，GV を用いる際には，スペクトル特徴量として，

Vol.2010-SLP-84 No.29
2010/12/21

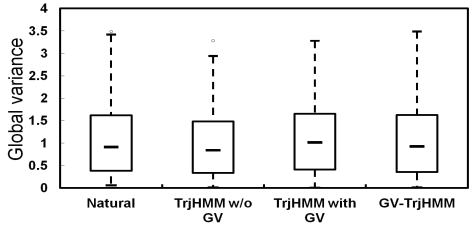

図 9 100 フレームまで (無音区間は除く) で計算された 1 次のメルケプストラム係数の GV の分布
Fig. 9 Distribution of GV of the 1st mel-cepstral coefficient calculated over 100 frames except for silence frames

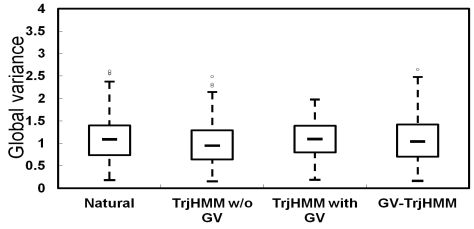

図 10 300 フレームまで (無音区間は除く) で計算された 1 次のメルケプストラム係数の GV の分布
Fig. 10 Distribution of GV of the 1st mel-cepstral coefficient calculated over 300 frames except for silence frames

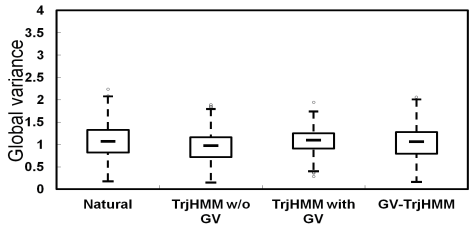

図 11 500 フレームまで (無音区間は除く) で計算された 1 次のメルケプストラム係数の GV の分布
Fig. 11 Distribution of GV of the 1st mel-cepstral coefficient calculated over 500 frames except for silence frames

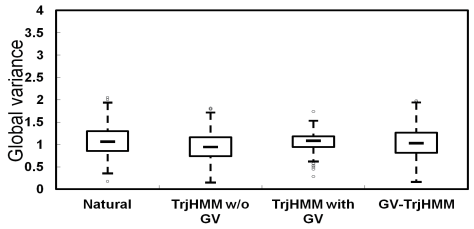

図 12 700 フレームまで (無音区間は除く) で計算された 1 次のメルケプストラム係数の GV の分布
Fig. 12 Distribution of GV of the 1st mel-cepstral coefficient calculated over 700 frames except for silence frames

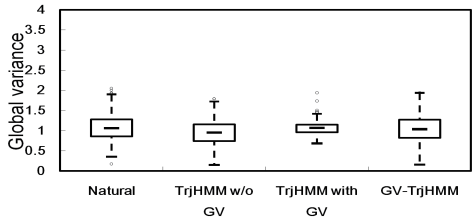

図 13 900 フレームまで (無音区間は除く) で計算された 1 次のメルケプストラム係数の GV の分布
Fig. 13 Distribution of GV of the 1st mel-cepstral coefficient calculated over 900 frames except for silence frames

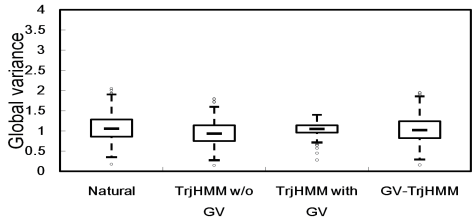

図 14 全てのフレームまで (無音区間は除く) で計算された 1 次のメルケプストラム係数の GV の分布
Fig. 14 Distribution of GV of the 1st mel-cepstral coefficient calculated over all frames except for silence frames

メル LSP よりもメルケプストラムの方が適していることを，主観評価実験結果から確認した．今後，フレーム数を考慮した GV のモデリング法について検討する予定である．

参 考 文 献

1) Y. Sagisaka. “Speech synthesis by rule using an optimal election of non–uniform synthesis units,” *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 679–682, 1988.
2) H. Zen, K. Tokuda, A. Black. “Statistical parametric speech synthesis,” *Speech Communication*, Vol. 51, pp 1039–1064, 2009.
3) 徳田 恵一，益子 貴史，小林 隆夫，今井 聖．“動的特徴を用いた HMM からの音声パラメータ生成アルゴリズム,” 音響誌, Vol. 53, no. 3, pp. 192–200, 1997.
4) Y. Qian, Z. Yan, Y. Wu, F. Soong, G. Zhang, and L. Wang. “An HMM Trajectory Tiling (HTT) Approach to High Quality TTS -Microsoft Entry to Blizzard Challenge 2010,” *Proceedings of Blizzard Challenge Workshop*, 2010.
5) Y. Wu, and R. Wang. “Minimum generation error training for HMM–based speech synthesis,” *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 14–19, 2006.
6) H. Zen, K. Tokuda, and T. kitamura. “Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences,” *Computer Speech and Language*, Vol. 21, pp. 153–173, 2007.
7) T. Toda, A. W. Black, and K. Tokuda. “Voice conversion based on maximum– likelihood estimation of spectral parameter trajectory,” *IEEE Trans. Audio speech and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.
8) T. Toda, and K. Tokuda. “A speeech parameter generation algorithm considering global variance for HMM–based speech synthesis,” *IEICE Transactions*, Vol. E90–D, No. 5, pp. 816–824, 2007.
9) T. Toda, and S. Young. “Trajectory training considering global variance for HMM–based speech synthesis,” *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4025–4028, 2009.
10) H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. “A Hidden Semi– Markov Model–Based Speech Synthesis System,” *IEICE Transactions*, Vol. E90– D, No. 5, pp. 825–834, 2007.
11) H. Zen, T. Toda, and K. Tokuda. “The Nitech–NAIST HMM–based speech synthesis system for the Blizzard Challenge 2006,” *IEICE Transactions*, Vol. E91–D, No. 6, pp. 1764–1773, 2008.
12) H. Kawahara, I. Katsuse, and A. Cheveigne. “Restructuring speech representations using a pitch–adaptive time–frequency smoothing and an instantaneous–frequency– based F0 extraction: possible role of a repetitive structure in sounds,” *Speech Communication*, Vol. 27, No. 3–4, pp. 187–207, 1999.
13) K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi. “Multi-Space Probability Distribution HMM,” *IEICE Transactions*, Vol. E85–D, No. 3, pp. 455–464, 2002.
14) K. Shinoda and T. Watanabe. “MDL-based context-dependent subword modeling for speech recognition,” *J. Acoust. Soc. Jpn. (E)*, Vol. 21, No. 2, pp. 79–86, 2000.
15) 大谷 大和，戸田 智基，猿渡 洋，鹿野 清宏．“STRAIGHT 混合励振源を用いた混合正規分布モデルに基づく最ゆう声質変換法,” 信学論，Vol. J91–D, No. 4, pp. 1082–1091, 2008.
16) 今井 聖，住田 一男，古市 千枝子．“音声合成のためのメル対数スペクトル近似 (MLSA) フィルタ,” 信学論，Vol. J66–A, No. 2, pp. 122–129, 1983.