Vol.2019-SLP-126 No.1
2019/2/27

GMMNに基づく音声合成における グラム行列のスパース近似の検討

郡山 知樹[1,a)]　高道 慎之介[2,b)]　小林 隆夫[1,c)]

概要：人間の音声生成のように発話間変動を持つ音声合成の実現を目標とし，我々はこれまでに，生成的モーメントマッチングネットワーク (GMNN) に基づく音声パラメータのランダム生成手法を提案している．GMMN では分布間の距離を表す条件付き maximum mean discrepancy(CMMD) を最小にするようにニューラルネットワークを学習する．音声合成のように学習データのサイズが大きい場合，CMMD を直接求めることは計算量の観点から非現実的であり何らかの近似を行う必要があったが，これまで近似手法について十分な検討が行われていなかった．本研究では CMMD の計算手法として，変数同士の類似度を表すグラム行列に random Fourier features (RFF) を用いる近似手法を提案し従来のブロック対角近似手法との比較を行う．またミニバッチの選択手法して，従来のランダム選択の代わりに K-means クラスタリングを用いて，類似した入力変数を同じミニバッチとする手法を検討する．主観評価実験では提案法が従来法に比べ，発話間変動が知覚されやすいという結果を得た．

キーワード：生成モーメントマッチングネットワーク, 統計的音声合成, ニューラルネットワーク, maximum mean discrepancy，カーネル法

A Study of Sparse Approximation of Gram Matrices for GMMN-based Speech Synthesis

Abstract: To realize human-like synthetic speech, synthetic speech samples should change every time even if the same sentences is spoken. In this context, we have proposed a technique of random sampling of synthetic speech parameters based on generative moment matching network (GMMN). GMMN is a neural network whose parameters are trained using conditional maximum mean discrepancy (CMMD) which represents the distance of two distributions. An issue of GMMN is that CMMD is computationally infeasible for a large amount of data, including speech synthesis database. In this report, we propose an approximation method based on random Fourier features and minibatch selection technique using K-means clustering. In the subjective evaluations, the proposed method outperformed the conventional one in the perception of inter-speech diversity.

Keywords: Generative moment matching network, statistical speech synthesis, neural network, maximum mean discrepancy, kernel method

1. はじめに

深層ニューラルネット (DNN) に基づく音声合成は近年広く研究されており，自然音声と知覚的に差がない合成音声を生成できるという報告もある [1], [2]．しかし，多くの DNN に基づく音声合成モデルは言語情報から音声情報への一対一の写像を学習しているため，同じ文に対して毎回知覚的に変化のない音声を生成してしまう．それに対し，人間が同じ文を発する場合，その音声には発話間変動があり，必ずしも同じように話すとは限らない．したがって，このような現象を音声合成システムにおいて再現するためには，同じ入力に対しどのように発話がゆらぐかをモデル

[1] 東京工業大学 工学院 情報通信系
G2-4, Nagatsuta-cho, Midori-ku, Yokohama, 226–8502, Japan.
[2] 東京大学 大学院情報理工学研究科 システム情報学専攻
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan.
a) koriyama@ip.titech.ac.jp
b) shinnosuke.takamichi@ipc.i.u-tokyo-ac.jp
c) takao.kobayashi@ip.titech.ac.jp

化する必要がある．

これに対し我々は生成的モーメントマッチングネットワーク (GMMN) に基づく音響特徴量のランダムサンプリング手法を提案している [3]．GMMN は入力変数に対する出力変数の分布を予測する生成モデルであり，入力変数と乱数を組み合わせることで出力分布のサンプル点を分布の形状を考慮せず直接求めることが可能である．GMMN に基づく音声合成では，入力変数を言語情報から得られるコンテキストに，出力変数をメルケプストラムや基本周波数などの音声パラメータとすることで，同じコンテキストであっても乱数の値によって異なる音声パラメータを出力することを可能にしている．また，我々はこれまでに話者照合に用いる i-vector の生成 [4] や歌声合成におけるダブルトラッキングの生成 [5] に GMMN を適用し，その有効性を確認している．

一方で，GMMN の問題点として学習時における計算コストが挙げられる．学習データのフレーム数が N のとき，GMMN の学習基準である CMMD（条件付き maximum mean discrepancy）は $\mathcal{O}(N^3)$ の計算量を必要とする．具体的には，入力変数に対するグラム行列の逆行列の計算 ($\mathcal{O}(N^3)$)，および出力変数のグラム行列の計算 ($\mathcal{O}(N^2)$) がボトルネックとなっている．そのため，音声合成において CMMD を直接計算することは現実的ではない．これまでの検討 [3] では，ランダムに選択したミニバッチそれぞれに対して CMMD を計算し，その和を学習時の目的関数としていた．これは入力変数および出力変数のグラム行列をブロック対角行列で近似することに一致するが，近似手法によって性能が変わることが予想される．

そのため，本研究では近似手法が GMMN に基づく音声合成に与える影響を調査する．近似手法としては，データの一部分だけに注目するローカルな近似手法であるブロック対角に加え，学習データ全体のグラム行列を低ランク行列で近似する random Fourier features (RFF)[6] を用いる手法を提案する．また，ミニバッチを選ぶ際にランダムにフレームを選択するのではなく，入力変数の K-means クラスタリングによって類似したフレーム同士のコンテキストを選択する手法を提案する．実験では，合成音声の自然性だけでなく，同じ文からランダムに生成された 2 つの音声に対して違いを知覚するかを評価する実験を行い，近似手法およびミニバッチの選択法についてその有効性を検討する．

2. Maximum mean discrepancy による分布間距離

2.1 MMD

本稿ではまずモデルの学習基準である maximum mean discrepancy (MMD)[7] について説明する．MMD は空間 $\mathcal{Y}$ 上の確率変数を持つ 2 つの分布 P と $\tilde{P}$ の距離を表す指標であり，以下の式で定義される．

$$\mathrm{MMD} = \sup_{\|f\|\leq 1, f\in\mathcal{F}} \left| \mathbb{E}_{Y\sim P}[f(Y)] - \mathbb{E}_{\tilde{Y}\sim\tilde{P}}[f(\tilde{Y})] \right| \quad (1)$$

ただし，$\mathcal{F}$ は $\mathcal{Y}$ 上の正定値カーネル k の定める再生核ヒルベルト空間 (reproducing kernel Hilbert space, RKHS) とする．式 (1) は関数 f によって写像された値の平均値の差分の最大値を示し，この平均の差分により分布間の距離を表す．$\mathcal{Y}$ から $\mathcal{F}$ への写像を ϕ とすると，式 (1) の右辺第一項はカーネル平均埋め込み μ_Y を用いて，

$$\begin{aligned}\mathbb{E}_{Y\sim P}[f(Y)] &= \mathbb{E}_{Y\sim P}\langle f, \phi(Y)\rangle_{\mathcal{F}} \\ &= \langle f, \mathbb{E}_{Y\sim P}[\phi(Y)]\rangle_{\mathcal{F}} = \langle f, \mu_Y\rangle_{\mathcal{F}} \end{aligned} \quad (2)$$

となる．ここで $\langle\cdot,\cdot\rangle_{\mathcal{F}}$ は空間 $\mathcal{F}$ 上の内積を表す．これを用いると，

$$\mathrm{MMD} = \sup_{\|f\|\leq 1, f\in\mathcal{F}} \langle f, \mu_Y - \mu_{\tilde{Y}}\rangle_{\mathcal{F}} \quad (3)$$

となり．これを最大にする f は

$$f = \frac{\mu_Y - \mu_{\tilde{Y}}}{\|\mu_Y - \mu_{\tilde{Y}}\|_{\mathcal{F}}} \quad (4)$$

で与えられるため，式 (1) は結果として

$$\begin{aligned}\mathrm{MMD}^2 &= \|\mu_Y - \mu_{\tilde{Y}}\|^2_{\mathcal{F}} \\ &= \langle\mu_Y, \mu_Y\rangle_{\mathcal{F}} + \langle\mu_{\tilde{Y}}, \mu_{\tilde{Y}}\rangle_{\mathcal{F}} - 2\langle\mu_Y, \mu_{\tilde{Y}}\rangle_{\mathcal{F}} \end{aligned} \quad (5)$$

で表される．

分布 P のサンプルが $\mathcal{D}_{\mathbf{Y}} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ で得られるとき，行列表現を用いて $\mathbf{\Phi} = [\phi(\mathbf{y}_1), \ldots, \phi(\mathbf{y}_N)]^\top$ とすると，標本平均 $\hat{\mu}_Y$ は以下の式で表される．

$$\hat{\mu}_Y = \frac{1}{N}\mathbf{\Phi}_{\mathbf{Y}}^\top \mathbf{1}_{N\times 1} \quad (6)$$

ただし，$\mathbf{1}_{A\times B}$ は要素がすべて 1 の $(A\times B)$ 行列である．標本平均を用いると式 (5) の内積は，

$$\begin{aligned}\langle\mu_Y, \mu_{\tilde{Y}}\rangle_{\mathcal{F}} &= \frac{1}{N\tilde{N}} \mathrm{Tr}\left[(\mathbf{\Phi}_{\mathbf{Y}}^\top \mathbf{1}_{N\times 1})^\top (\mathbf{\Phi}_{\tilde{\mathbf{Y}}}^\top \mathbf{1}_{\tilde{N}\times 1})\right] \\ &= \frac{1}{N\tilde{N}} \mathrm{Tr}\left[\mathbf{K}_{\mathbf{Y}\tilde{\mathbf{Y}}} \mathbf{1}_{\tilde{N}\times N}\right] \end{aligned} \quad (7)$$

となる．だたし，$\mathbf{K}_{\mathbf{Y}\tilde{\mathbf{Y}}} = \mathbf{\Phi}_{\mathbf{Y}}\mathbf{\Phi}_{\tilde{\mathbf{Y}}}^\top$ は $\mathcal{D}_{\mathbf{Y}}$, $\mathcal{D}_{\tilde{\mathbf{Y}}}$ から得られるグラム行列である．グラム行列の要素はカーネル関数 $k(\cdot,\cdot)$ によって求められるため，写像 $\phi(\cdot)$ による無限次元ベクトルを直接求める必要がない．式 (5) の他の内積に対して同様の式変形を行うことで

$$\begin{aligned}\mathrm{MMD}^2 = &\frac{1}{N^2}\mathrm{Tr}\left[\mathbf{K}_{\mathbf{YY}}\mathbf{1}_{N\times N}\right] + \frac{1}{\tilde{N}^2}\mathrm{Tr}\left[\mathbf{K}_{\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}}\mathbf{1}_{\tilde{N}\times\tilde{N}}\right] \\ &- 2\frac{1}{N\tilde{N}}\mathrm{Tr}\left[\mathbf{K}_{\mathbf{Y}\tilde{\mathbf{Y}}}\mathbf{1}_{\tilde{N}\times N}\right] \end{aligned} \quad (8)$$

が得られる．MMD は上記のように無限次元のベクトルを考慮して，分布間の距離を求めることから，パラメトリックな分布を仮定しないノンパラメトリックな手法と呼ばれている．

2.2 条件付き MMD [8]

次に 2 つの分布 P，$\tilde{P}$ を，任意の入力ベクトル $\mathbf{x}(\in \mathcal{X})$ が与えられたときの確率変数 Y の条件付き分布とする場合を考える．このとき，MMD と同様に関数 $f \in \mathcal{F}$ を用いて分布間の距離を以下の式で表す．

$$\left|\mathbb{E}_{Y\sim P}[f(Y;\mathbf{x})] - \mathbb{E}_{\tilde{Y}\sim \tilde{P}}[f(\tilde{Y};\mathbf{x})]\right| \tag{9}$$

f が RKHS $\mathcal{F}$ の上の元であることから条件付き平均 $\mu_{Y|\mathbf{x}}$ を用いて

$$\mathbb{E}_{Y\sim P}[f(Y;\mathbf{x})] = \langle f, \mu_{Y|\mathbf{x}}\rangle_{\mathcal{F}} \tag{10}$$

と表せる．ここで，$\mathcal{F}$ と異なる RKHS $\mathcal{G}$ に対し，写像 $\psi : \mathcal{X} \to \mathcal{G}$ を考え，条件付き平均 $\mu_{Y|\mathbf{x}}$ が $\mathcal{G}$ 上の元 $\psi(\mathbf{x})$ からの線形変換，すなわち

$$\mu_{Y|\mathbf{x}} = C_{Y|X}\psi(\mathbf{x}) \tag{11}$$

で求められると仮定する．ただし $C_{Y|X}$ は条件付き共分散作用素と呼ばれる線形作用素である．このとき $C_{Y|X}$ はテンソル積空間 $\mathcal{F}\otimes\mathcal{G}$ 上の元であり，

$$\langle f, \mu_{Y|\mathbf{x}}\rangle_{\mathcal{F}} = \langle f, C_{Y|X}\psi(\mathbf{x})\rangle_{\mathcal{F}} \tag{12}$$

$$= \langle f\otimes\psi(\mathbf{x}), C_{Y|X}\rangle_{\mathcal{F}\otimes\mathcal{G}} \tag{13}$$

と変形できる [9]．ただし $\otimes$ はテンソル積である．したがって，式 (9) の条件付き分布間距離の最大値は $g(\mathbf{x}) = f\otimes\psi(\mathbf{x})$ とすると，

$$\mathrm{CMMD} = \sup_{\substack{\|g(\mathbf{x})\|\le 1\\ g(\mathbf{x})\in\mathcal{F}\otimes\mathcal{G}}} \langle g(\mathbf{x}), C_{Y|X} - C_{\tilde{Y}|\tilde{X}}\rangle_{\mathcal{F}\otimes\mathcal{G}} \tag{14}$$

となり，MMD と同様にして CMMD の定義

$$\mathrm{CMMD}^2 = \|C_{Y|X} - C_{\tilde{Y}|\tilde{X}}\|^2_{\mathcal{F}\otimes\mathcal{G}} \tag{15}$$

を得る．

データ $\mathcal{D} = \{(\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_1,\mathbf{y}_N)\}$ が与えられたとき，線形作用素 $C_{Y|X}$ の推定値は次の式で求められることが知られている [9]．

$$\hat{C}_{Y|X} = \mathbf{\Phi_Y}(\mathbf{\Upsilon_X}\mathbf{\Upsilon_X}^\top + \lambda\mathbf{I})^{-1}\mathbf{\Upsilon_X}^\top \tag{16}$$

ただし，$\mathbf{\Phi_Y} = [\phi(\mathbf{y}_1),\ldots,\phi(\mathbf{y}_N)]^\top$，$\mathbf{\Upsilon_X} = [\psi(\mathbf{x}_1),\ldots,\psi(\mathbf{x}_N)]^\top$ であり，$\lambda(>0)$ は正則化定数である．これを用いると CMMD の推定量は

$$\begin{aligned}\mathrm{CMMD}^2 = &\Big\|\mathbf{\Phi_Y}(\mathbf{\Upsilon_X}\mathbf{\Upsilon_X}^\top + \lambda\mathbf{I})^{-1}\mathbf{\Upsilon_X}^\top\\ &\quad - \mathbf{\Phi}_{\tilde{\mathbf{Y}}}(\mathbf{\Upsilon}_{\tilde{\mathbf{X}}}\mathbf{\Upsilon}_{\tilde{\mathbf{X}}}^\top + \lambda\mathbf{I})^{-1}\mathbf{\Upsilon}_{\tilde{\mathbf{X}}}^\top\Big\|^2_{\mathcal{F}\otimes\mathcal{G}}\\ = &\,\mathrm{Tr}\left[\mathbf{K}_{\mathbf{Y},\mathbf{Y}}\bar{\mathbf{H}}^{-1}_{\mathbf{X},\mathbf{X}}\mathbf{H}_{\mathbf{X},\mathbf{X}}\bar{\mathbf{H}}^{-1}_{\mathbf{X},\mathbf{X}}\right]\\ &+ \mathrm{Tr}\left[\mathbf{K}_{\tilde{\mathbf{Y}},\tilde{\mathbf{Y}}}\bar{\mathbf{H}}^{-1}_{\tilde{\mathbf{X}},\tilde{\mathbf{X}}}\mathbf{H}_{\tilde{\mathbf{X}},\tilde{\mathbf{X}}}\bar{\mathbf{H}}^{-1}_{\tilde{\mathbf{X}},\tilde{\mathbf{X}}}\right]\\ &- 2\,\mathrm{Tr}\left[\mathbf{K}_{\mathbf{Y},\tilde{\mathbf{Y}}}\bar{\mathbf{H}}^{-1}_{\tilde{\mathbf{X}},\tilde{\mathbf{X}}}\mathbf{H}_{\tilde{\mathbf{X}},\mathbf{X}}\bar{\mathbf{H}}^{-1}_{\mathbf{X},\mathbf{X}}\right]\end{aligned} \tag{17}$$

となる．ただし，$\mathbf{H}_{\mathbf{X},\tilde{\mathbf{X}}} = \mathbf{\Upsilon_X}\mathbf{\Upsilon}_{\tilde{\mathbf{X}}}^\top$ は入力変数に対するグラム行列，$\bar{\mathbf{H}}_{\mathbf{X},\tilde{\mathbf{X}}} = \mathbf{H}_{\mathbf{X},\tilde{\mathbf{X}}} + \lambda\mathbf{I}$ であり，他の添え字についても同様である．

ここで説明を簡単にするため，2 つデータで入力 $(\mathbf{x}_1,\ldots,\mathbf{x}_N)$ が共通である場合を考える．このとき，$\mathbf{H} \triangleq \mathbf{H}_{\mathbf{X},\mathbf{X}} = \mathbf{H}_{\mathbf{X},\tilde{\mathbf{X}}} = \mathbf{H}_{\tilde{\mathbf{X}},\tilde{\mathbf{X}}}$ となり，CMMD は以下の式のように簡略化できる．

$$\mathrm{CMMD}^2 = \mathrm{Tr}\left[\left(\mathbf{K}_{\mathbf{Y},\mathbf{Y}} + \mathbf{K}_{\tilde{\mathbf{Y}},\tilde{\mathbf{Y}}} - 2\mathbf{K}_{\mathbf{Y},\tilde{\mathbf{Y}}}\right)\mathbf{L}\right] \tag{18}$$

$$\mathbf{L} = (\mathbf{H}+\lambda\mathbf{I})^{-1}\mathbf{H}(\mathbf{H}+\lambda\mathbf{I})^{-1} \tag{19}$$

MMD と CMMD の違いは，MMD では出力変数 Y のグラム行列 $\mathbf{K}$ に掛けられる行列が壱行列であるのに対し，CMMD では入力変数から得られる行列 $\mathbf{L}$ が使用されている点である．行列 $\mathbf{L}$ の要素は入力変数の値が近いほど大きい値を持つ傾向があり，これによって CMMD は入力変数の値によって重み付けした MMD と見なすことができる．

3. GMNN に基づく音声合成

GMMN に基づく音声合成 [3] では，フレームレベルのコンテキストが与えられたときの音響特徴量の条件付き分布を，ニューラルネットにより予測する．**図 1** に本研究で用いるネットワーク構造を示す．ネットワークは，平均二乗誤差 (MSE) 基準で学習を行う DNN と，DNN のボトルネック特徴量 $\mathbf{e}$ と乱数 $\mathbf{n}$ の結合ベクトルを入力として発話間変動を出力する GMMN で構成される．学習時にはまず MSE 基準の DNN を行う．次に得られたボトルネック特徴量 $\mathbf{e}$ を用いて GMMN を学習にする．このとき，MSE 基準の DNN のパラメータを固定した上で CMMD を誤差関数として使用し，誤差逆伝搬法によりパラメータを学習する．

なお，フレームレベルのコンテキストは高次元であるため，カーネル関数が疎になりやすいという問題点がある．そこで，ボトルネック特徴量を GMMN の入力としている．また，GMNN の出力は MSE 基準 DNN との差分とする．GMNN の出力は必ずしも差分である必要はなく音響特徴量を直接予測することが可能ではあるものの，事前実験において学習が安定しなかったため，本研究では差分を予測するモデルを用いる．

4. GMNN 学習のための CMMD の計算手法

CMMD を計算するにあたり，学習データの総フレーム数を N としたとき，式 (19) において，逆行列の計算には $\mathcal{O}(N^3)$ の計算量が必要となる．またトレースの計算には $\mathcal{O}(N^2)$ の計算量を必要とする．そのため，N が大きい場合 CMMD 自体の計算が困難であり，さらに確率的勾配法に基づくパラメータ最適化は現実的でない．そこで本研究では，目的関数である CMMD をミニバッチごとの値の和

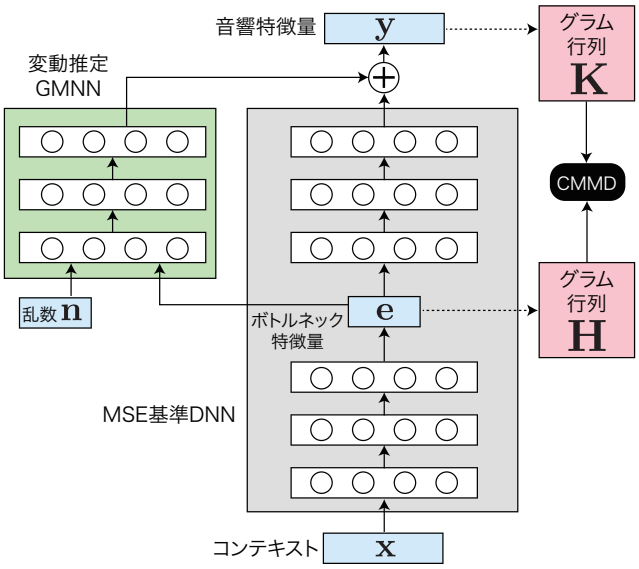

図 1 GMMN に基づく音声合成のネットワーク図
Fig. 1 Network of GMMN-based speech synthesis.

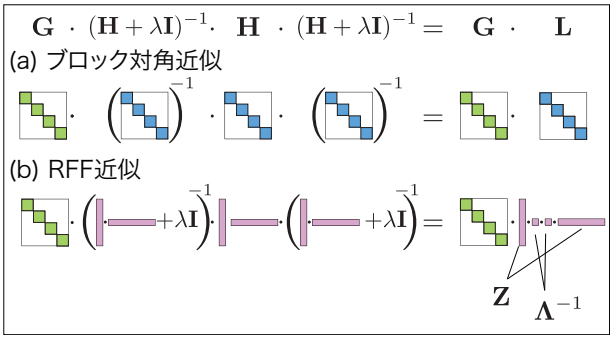

図 2 入力変数のグラム行列の近似手法の比較
Fig. 2 Comparison of approximation method of Gram matrices for input variables.

の形式で近似する手法を検討する．

4.1 グラム行列の近似

まず，出力変数に対するグラム行列に対応する行列 $\mathbf{G} \triangleq (\mathbf{K}_{\mathbf{Y}} + \mathbf{K}_{\tilde{\mathbf{Y}}} - 2\mathbf{K}_{\mathbf{y},\tilde{\mathbf{Y}}})$ をミニバッチに従って $\mathrm{diag}[\mathbf{G}_1, \ldots, \mathbf{G}_S]$ となるようにブロック対角行列で近似する．$\mathbf{L}$ のブロック対角成分を $\mathbf{L}_i$ $(i = 1 \ldots S)$ としたとき，CMMD は $\mathcal{L} = \sum_{i=1}^{S} \mathrm{Tr}[\mathbf{G}_i \mathbf{L}_i]$ で近似されるため，確率的勾配法によるパラメータ最適化が可能になる．次に $\mathbf{L}$ のブロック対角成分を求める手法として，従来法である $\mathbf{H}$ をブロック対角行列で近似する手法に加え，random Fourier features (RFF) [6] に基づく低ランク近似を用いる手法を提案する．この 2 手法の近似方法の違いを**図 2** に示す．

入力変数のグラム行列 $\mathbf{H}$ を $\mathrm{diag}[\mathbf{H}_1, \ldots, \mathbf{H}_S]$ のようにブロック対角近似する場合，$\mathbf{L}$ も $\mathrm{diag}[\mathbf{L}_1, \ldots, \mathbf{L}_S]$ で表されるブロック対角行列となり，その値は $\mathbf{L}_i = (\mathbf{H}_i + \lambda\mathbf{I})^{-1}\mathbf{H}_i(\mathbf{H}_i + \lambda\mathbf{I})^{-1}$ で求められる．このとき逆行列の計算はミニバッチごとに行うため，$\mathbf{L}_i$ の計算量はミニバッチサイズ B に対して $\mathcal{O}(B^3)$ となる．この手法はこれまでの研究 [3] で行った，ミニバッチごとの CMMD を求めてパラメータの更新を行う手法と等価である．

RFF[6] はカーネル関数を，有限次元のベクトルの内積で近似する手法である．例えば広く使用される RBF カーネルは以下の式で表される．

$$k_{\mathrm{RBF}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|^2\right) \tag{20}$$

この関数は正規乱数 $\boldsymbol{\omega}_r \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ と一様乱数 $b_r \sim \mathcal{U}[0, 2\pi)$ を用いてモンテカルロ近似ができることが知られている．

$$k_{\mathrm{RBF}}(\mathbf{x}, \mathbf{x}') \approx \frac{1}{M}\sum_{r=1}^{M} \cos(\mathbf{x}^\top \boldsymbol{\omega}_r + b_r)\cos(\mathbf{x}'^\top \boldsymbol{\omega}_r + b_r) \tag{21}$$

ここで M は RFF の次元であり，$M(\ll N)$ 次元の RFF から成る $(N \times M)$ の入力変数行列

$$\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]^\top \tag{22}$$

$$\mathbf{z}_n = \left[\cos(\mathbf{x}_n^\top \boldsymbol{\omega}_1 + b_1), \ldots, \cos(\mathbf{x}_n^\top \boldsymbol{\omega}_M + b_M)\right]^\top \tag{23}$$

を使用すると，グラム行列 $\mathbf{H}$ は $\mathbf{Z}\mathbf{Z}^\top$ で表される低ランク行列で近似できる．ここで Woodbury の公式を用いると，行列 $\mathbf{L}$ は

$$\begin{aligned}\mathbf{L} &\approx (\mathbf{Z}\mathbf{Z}^\top + \lambda\mathbf{I})^{-1}\mathbf{Z}\mathbf{Z}^\top(\mathbf{Z}\mathbf{Z}^\top + \lambda\mathbf{I})^{-1} \\ &= \mathbf{Z}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}^{-1}\mathbf{Z}^\top \end{aligned} \tag{24}$$

$$\boldsymbol{\Lambda} = \mathbf{Z}^\top\mathbf{Z}/\lambda + \mathbf{I} \tag{25}$$

で近似できる．$\boldsymbol{\Lambda}$ は GMMN のパラメータに依存しないため，あらかじめ計算しておくことができる．このとき，ブロック対角成分 $\mathbf{L}_i$ は $\mathcal{O}(BM^2)$ の計算量で求められる．

上記 2 手法を比較すると，$\mathbf{H}$ のブロック対角近似を用いる手法では，ミニバッチごとにローカルな線形作用素 $\mathbf{C}_{Y|X}$ を求めてその差分を計算しているのに対し，RFF 近似ではデータ全体から求められるグローバルな線形作用素 $\mathbf{C}_{Y|X}$ の差分を計算していると考えることができる．また，RFF 近似では $\mathcal{O}(B^3)$ の逆行列計算を必要としないため，バッチサイズを大きくしても計算時間が大きく増加しないというメリットがある．

4.2 クラスタリングに基づくミニバッチ選択

本研究ではさらに，ミニバッチの選択方法を検討する．出力変数のグラム行列 $\mathbf{K}$ は音響特徴量の多様性からスパースになりやすく，例えば異なる音素間ではカーネル関数の値が非常に小さくなり，ネットワークに伝播する勾配も小さくなると考えられる．そこで，クラスタリングによって類似したコンテキストを同じミニバッチに割り当てることで，$\mathbf{K}$ のブロック対角近似の影響が小さくなると予想され

る．クラスタリングには GMMN の入力であるボトルネック特徴量 $\mathbf{e}$ に対して，各クラスタのサイズが既定値を下回るまで，再帰的に 2 クラス K-means による分割を繰り返す手法を用いる．

5. 実験

5.1 実験条件

実験には日本語女性話者による異なる 203 文をそれぞれ 5 回読み上げた繰り返し発話を用いた．このデータの一部は JSUT[10] コーパスに含まれている．このうち 150 文を学習セット，26 文を開発セット，27 文をテストセットとした．16kHz にダウンサンプリングした音声信号から，5ms 毎に STRAIGHT を用いて f_0，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_0，5 次元の非周期性指標，およびそれらの Δ, Δ^2 と有声／無声フラグから成る 139 次元ベクトルを音響特徴量として使用した．入力変数であるフレームレベルのコンテキストには 556 次元のベクトルを用いた．入力ベクトルは平均 0，分散 1 となるように，出力ベクトルは値が $[-1, 1]$ の範囲になるように，それぞれ正規化を行った．

従来法 [3] と同様に GMMN の入力にはコンテキストではなくボトルネック特徴量を用いた．ボトルネック特徴量の次元は 128，GMMN の乱数の次元は 3 とした．モデル構造として，ボトルネック特徴量を抽出する DNN にはエンコーダ部分，デコーダ部分がそれぞれ隠れ層 3 層，素子数 512 のネットワークを用いた．また，GMMN の構造は隠れ層 3 層，素子数 512 の DNN とし，ボトルネック特徴量抽出 DNN の出力との差分をネットワークの出力とした．それぞれのネットワークにおいて，隠れ層の活性化関数には ReLU，出力およびボトルネック層の活性化関数には tanh を用いた．欠落率 20%のドロップアウト，係数 10^{-6} の重み減衰による正則化を行い，バッチ正規化による勾配消失対策を行った．最適化には Adam を使用し学習率を 0.001 として最大 300 エポックの繰り返しを行った．ボトルネック特徴量を抽出する DNN の学習におけるミニバッチサイズは 1024 とした．GMMN 学習におけるミニバッチには，ランダム選択の場合に 10000 サンプル，クラスタリングの場合は最大 10000 サンプルとなるように分割した．

CMMD のカーネル関数には $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/2\sigma^2)$ で表される RBF カーネルを用いた．パラメータ σ の値として，入力変数のカーネルに対しては入力ベクトル同士のユークリッド距離の最大値の半分の値を，出力変数に対してはユークリッド距離の中央値を用いた．

5.2 主観評価結果

聴取実験では，合成音声の自然性に加え，同じ文章に対しランダムに生成された 2 つの音声を聴いて発話間に変動が知覚できるかを評価する試験を行った．合成音声は各文章に対し 5 パターンランダム生成を行った．聴取者は 60 名で，各聴取者はランダムに選ばれた 3 文章を 5 段階のスコアで評価した．自然性および発話間変動の結果をそれぞれ**図 3**，**図 4** に示す．図中の MSE は平均二乗誤差基準で学習したランダム生成の行われない DNN，VOC は上限である分析合成音である．LOCAL はブロック対角近似，RFF は RFF 近似をそれぞれ表し，ミニバッチをランダムに選択した場合を RAND，クラスタリングを行った場合を CLST とした．

発話間変動の評価では RFF-CLST が他の手法に比べて発話間のばらつきが知覚できるという結果となった．一方で自然性の評価では RFF-CLST は RFF-RAND に比べ自

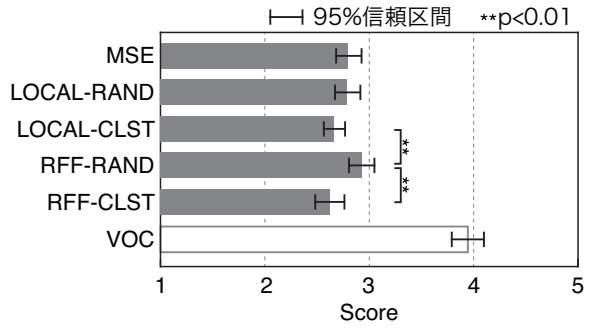

図 3 自然性の主観評価結果（1: とても低い音質，5: とても高い音質）

Fig. 3 Subjective evaluation on naturalness (1: too bad, 5: very good).

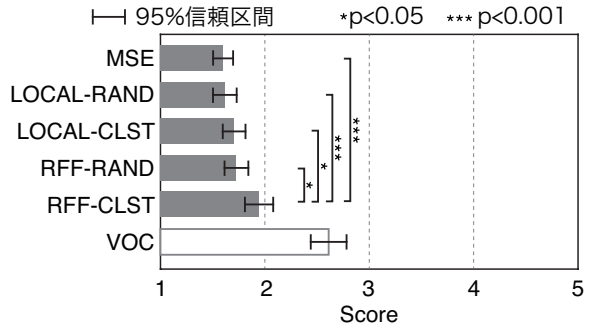

図 4 発話間変動の主観評価結果（1: 全く同じ，5: かなり違う）

Fig. 4 Subjective evaluation on diversity of two synthetic speech samples (1: completely equivalent, 5: very different).

表 1 音声合成パラメータのフレーム内標準偏差の平均

Table 1 Average of frame-level standard deviation of synthetic speech parameters.

	メルケプストラム		対数 f_0	音素継続長
	0th	1st	[cent]	[msec]
LOCAL-RAND	0.0230	0.0121	15.80	2.46
LOCAL-CLST	0.0528	0.0216	18.23	3.50
RFF-RAND	0.0208	0.0066	1.54	3.77
RFF-CLST	0.0493	0.0266	13.97	5.47

然性が有意に低いという結果となった．この理由としては発話間のばらつきが知覚できる代わりに，不自然に感じる音声になっていたという理由が考えられる．

5.3 発話間変動の客観評価

主観評価実験の結果について考察を行うため，発話間の変動の客観評価を行った．ランダムに 5 パターン生成された音響特徴量に対し，フレーム毎（音素継続長の場合は音素毎）に標準偏差を計算し，全フレームに対する平均を求めた．

メルケプストラムの 0 次および 1 次，対数 f_0，音素継続長に対する結果を**表 1** に示す．メルケプトラムではクラスタリングを行う LOCAL-CLST や RFF-CLST の標準偏差が大きくなった．一方で音素継続長では RFF-CLST が最も標準偏差が大きく RFF-RAND が続いている．一方で RFF-RAND は対数 f_0 においてほとんどばらつきが見られなかった．また，主観評価結果と比較すると，主観評価における音素継続長の標準偏差が大きいほど図 4 の発話間変動の主観評価結果のスコアが高くなる傾向にあることがわかる．したがって，このような主観評価結果は音素継続長のばらつきの影響が大きいと考えられる．

6. おわりに

本稿では，GMMN に基づく音声合成において，目的関数である CMMD の近似方法を検討した．入力変数のグラム行列の近似手法としてブロック対角近似に加え，RFF に基づく近似手法を提案した．また，ランダムなミニバッチ選択ではなく，K-means クラスタリングに基づくミニバッチ選択法を提案した．実験では RFF 近似かつクラスタリングに基づくミニバッチの選択法を用いることで，発話間変動が知覚されやすくなることを示した．今後の課題として，フレームレベルではなく系列レベルのモデリングの検討が挙げられる．

謝辞 本研究の一部は，セコム科学技術支援財団の助成を受け実施した．

参考文献

[1] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R. et al.: Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions, *Proc. ICASSP*, pp. 4779–4783 (2018).

[2] Luong, H.-T., Wang, X., Yamagishi, J. and Nishizawa, N.: Investigating Accuracy of Pitch-accent Annotations in Neural Network-based Speech Synthesis and Denoising Effects, *Proc. INTERSPEECH*, pp. 37–41 (2018).

[3] Takamichi, S., Koriyama, T. and Saruwatari, H.: Sampling-Based Speech Parameter Generation Using Moment-Matching Networks, *Proc. INTERSPEECH*, pp. 3961–3965 (2017).

[4] Shiota, S., Takamichi, S., and Matsui, T.: Data augmentation with moment-matching networks for i-vector based speaker verification, pp. 345–349 (2018).

[5] Tamaru, H., Saito, Y., Takamichi, S., Koriyama, T. and Saruwatari, H.: Generative Moment Matching Network-based Random Modulation Post-filter For DNN-based Singing Voice Synthesis And Neural Double-tracking, *Proc. ICASSP (accepted)* (2019).

[6] Rahimi, A. and Recht, B.: Random features for large-scale kernel machines, *Proc. NIPS*, pp. 1177–1184 (2008).

[7] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A.: A kernel two-sample test, *Journal of Machine Learning Research*, Vol. 13, pp. 723–773 (2012).

[8] Ren, Y., Zhu, J., Li, J. and Luo, Y.: Conditional generative moment-matching networks, *Proc. NIPS*, pp. 2928–2936 (2016).

[9] Song, L., Huang, J., Smola, A. and Fukumizu, K.: Hilbert space embeddings of conditional distributions with applications to dynamical systems, *Proc. ICML*, pp. 961–968 (2009).

[10] Sonobe, R., Takamichi, S. and Saruwatari, H.: JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis, *arXiv preprint arXiv:1711.00354* (2017).