社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

音素事後確率と *d*-vector を用いた Variational Autoencoder によるノンパラレル多対多音声変換

齋藤　佑樹[†,††]　井島　勇祐[†]　西田　京介[†]　高道慎之介[††]

† 日本電信電話株式会社　〒 239-0847 神奈川県横須賀市光の丘 1-1
†† 東京大学大学院 情報理工学系研究科　〒 113-8656 東京都文京区本郷 7-3-1

あらまし　話者コードで条件付けされた Variational AutoEncoder (VAE) を用いた従来のノンパラレル音声変換では，発話内容を表す潜在変数の過剰な正則化により，変換音声の品質が著しく劣化する．これに対し，本稿では，話者コードのみならず，学習済みの音声認識モデルの予測結果として得られる音素事後確率で条件付けされた VAE の学習法を提案する．本稿ではさらに，一対一 VAE 音声変換を任意話者対での変換が可能な多対多音声変換に拡張するための手法として，(1) 話者コードの適応，及び (2) 話者認証において有効な *d*-vector を用いた学習・変換法を比較する．実験的評価により，(1) 音素事後確率の導入により変換音声の品質が劇的に改善すること，及び (2) 話者コードと *d*-vector の両方がノンパラレル多対多 VAE 音声変換に適用可能であることを示す．

キーワード　ノンパラレル音声変換，多対多音声変換，variational autoencoder，音素事後確率，*d*-vector

Non-parallel and Many-to-Many Voice Conversion Using Variational Autoencoder Conditioned by Phonetic Posteriorgrams and *d*-vectors

Yuki SAITO[†,††], Yusuke IJIMA[†], Kyosuke NISHIDA[†], and Shinnosuke TAKAMICHI[††]

† Nippon Telegraph and Telephone Corporation Hikarinooka 1-1, Yokosuka-shi, Kanagawa, 239-0847 Japan
†† Graduate School of Information Science and Technology, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-8656 Japan

Abstract　This paper proposes novel frameworks for non-parallel and many-to-many voice conversion (VC) using variational autoencoders (VAEs). In conventional VAE-based VC, converted speech quality is significantly degraded due to an over-regularization of latent variables representing phonetic contents. To overcome the issue, this paper proposes a VAE-based non-parallel VC conditioned by not only the speaker codes but also phonetic posteriorgrams (PPGs) predicted from pre-trained speech recognition models. This paper also extends the conventional VC to many-to-many VC that can convert arbitrary speakers' characteristics into another ones. We compare two methods to realize this: 1) speaker code adaptation, and 2) the use of *d*-vectors obtained by using pre-trained speaker verification models. Experimental results demonstrate that 1) PPGs successfully improve converted speech quality, and 2) both speaker codes and *d*-vectors can be adopted to the VAE-based non-parallel and many-to-many VC.

Key words　Non-parallel voice conversion, many-to-many voice conversion, variational autoencoders, phonetic posteriorgrams, *d*-vectors

1. はじめに

音声変換 [1] とは，入力された音声の言語情報を保持しつつ，パラ言語・非言語情報を変換する技術である．近年，Deep Neural Network (DNN) を用いた音声変換 [2] が提案され，従来の Gaussian Mixture Model (GMM) [3] と比較して高精度な音声パラメータの変換が可能となっている．変換元・変換先話者での同一発話内容を収録したパラレルデータを用いて学習された DNN の音声変換モデルは，音声パラメータをフレーム毎に変換するため，高品質な音声変換を実現できる．しかし，パラレルデータの収集は困難であることが多いため，近年では，音声変換モデルの学習にパラレルデータを必要としないノンパラレル音声変換の手法が研究されている．特に，Variational AutoEncoder (VAE) [4] を用いたノンパラレル音声変換 [5] は，

従来の restricted Boltzmann machine を用いた手法 [6] と比較して学習が容易であることから，広く研究され始めている．

従来の VAE 音声変換 [5] では，encoder は入力された音声パラメータから話者非依存の潜在変数を教師なしに抽出し，decoder は潜在変数と与えられた話者表現を用いて音声パラメータを復元する．Encoder により抽出された潜在変数は，入力音声の発話内容を表すことが期待されるため [5]，decoder に入力する話者表現を変えることによって音声変換が実現される．しかし，潜在変数の分布が過度に単純化される過剰な正則化 [7] により，従来の VAE 音声変換の品質は，パラレルデータを用いて学習された DNN を用いた手法と比較して著しく劣化する．潜在変数の事前分布として GMM を用いる手法 [8] により，過剰な正則化の緩和が期待できるが，GMM の最適な混合数の決定は容易ではないため，VAE 音声変換への適用は困難であると予想される．

本稿では，VAE 音声変換の品質を改善させるために，話者表現のみならず，発話内容による条件付けを用いた VAE の学習法を提案する．提案手法では，話者非依存な音声認識モデルを構築するための比較的大規模なコーパスが利用可能であると仮定し，発話内容を表す潜在変数として，音声認識モデルの予測結果として得られる音素事後確率 [9] を導入する．提案手法における VAE の encoder は，入力された音声パラメータと音素事後確率から潜在変数を抽出し，decoder は，潜在変数，話者表現，そして音素事後確率を用いて音声パラメータを復元する．入力音声の発話内容は音素事後確率によって保持されるため，音韻の消失に起因する変換音声の品質劣化の改善が期待できる．本稿ではさらに，不特定多数の音声パラメータから話者非依存な潜在変数を抽出可能な VAE の性質に着目し，従来の VAE 音声変換を任意話者対での変換が可能な多対多音声変換へと拡張する．多対多音声変換における学習データに含まれない未知話者の話者表現を推定する手法として，本稿では (1) 従来の話者表現として用いられていた話者コード [10] の適応，及び (2) 話者認証モデルの中間表現として得られる d-vector [11] の利用を比較する．話者認証における d-vector の有効性はよく知られているため，任意話者の潜在変数を用いた学習・変換が可能である．実験的評価により，(1) 音素事後確率の導入により，VAE 音声変換の音質及び話者性が劇的に改善すること，及び (2) 話者コードと d-vector の両方が VAE を用いたノンパラレル多対多音声変換に適用可能であることを示す．

2. 従来の VAE 音声変換

2.1 話者コードによる条件付けを用いた VAE 音声変換 [5]

VAE の音声変換モデルは，音声パラメータ $\boldsymbol{x}$ が潜在変数 $\boldsymbol{z}$ と話者表現 $\boldsymbol{y}_{\mathrm{s}}$ から生成される確率モデルを表現する．従来の VAE 音声変換（図 1） [5] では，話者表現として one-hot ベクトルによる話者コード [10] を用いる．学習データに含まれる S 人のうちの i 番目の話者に対する話者コードは，次式で与えられる．

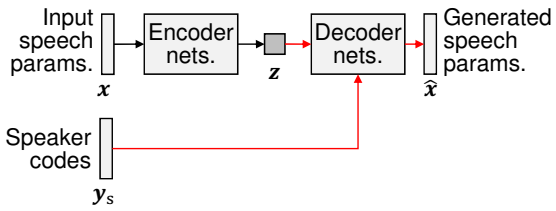

図 1　話者コードで条件付けされた VAE 音声変換の概略図．

Fig. 1　Conventional VAE-based VC conditioned by speaker codes.

$$y_{\mathrm{s}}^{(i)}(k) = \begin{cases} 1 & \text{if } k = i \\ 0 & \text{otherwise} \end{cases} \qquad (1 \leqq k \leqq S) \tag{1}$$

話者表現 $\boldsymbol{y}_{\mathrm{s}}$ と潜在変数 $\boldsymbol{z}$ が独立であると仮定し，VAE の学習では，$\boldsymbol{y}_{\mathrm{s}}$ で条件付けされた音声パラメータの周辺尤度 $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{y}_{\mathrm{s}}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{y}_{\mathrm{s}}) p_{\boldsymbol{\theta}}(\boldsymbol{z}) d\boldsymbol{z}$ を最大化するモデル（すなわち，decoder）パラメータ $\boldsymbol{\theta}$ を推定する．ここで，$p_{\boldsymbol{\theta}}(\boldsymbol{z})$ は潜在変数の事前分布である．周辺尤度に含まれる積分を直接計算することは困難であるため，VAE の学習では，話者非依存の encoder と，話者依存の decoder の 2 つのニューラルネットワークを導入する．encoder $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ は潜在変数の真の事後確率 $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ を近似し，decoder は音声パラメータの真の事後確率 $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{y}_{\mathrm{s}})$ を近似する．ここで，$\boldsymbol{\phi}$ は encoder のモデルパラメータである．VAE の学習時に最大化される対数尤度の変分下限は，次式で与えられる．

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x},\boldsymbol{y}_{\mathrm{s}}) = &- D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z})\right) \\ &+ \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{y}_{\mathrm{s}})\right] \end{aligned} \tag{2}$$

ここで，$D_{\mathrm{KL}}(\cdot||\cdot)$ は 2 つの確率分布間の Kullback-Leibler (KL) ダイバージェンスである．本稿では，式 (2) に含まれる KL ダイバージェンスを解析的に計算するために，encoder の事前分布 $p_{\boldsymbol{\theta}}(\boldsymbol{z})$ に多変量標準正規分布 $\mathcal{N}(\boldsymbol{z};\boldsymbol{0},\boldsymbol{I})$ を仮定する．Back-propagation 時の計算グラフの構築には，reparameterization trick [4] を用いる．図 2(a) に話者コードで条件付けされた VAE の有向グラフィカルモデルを示す．

学習後の VAE を用いた音声パラメータの変換時には，変換先話者の話者コードを decoder に入力する．例えば，入力された音声パラメータを学習データに含まれる j 番目の話者のものに変換するときには，$k = j$ のときに限り $y_{\mathrm{s}}^{(j)}(k) = 1$ となる話者コード $\boldsymbol{y}_{\mathrm{s}}^{(j)}$ を decoder に入力する．

2.2 従来手法の問題点

話者表現との独立性の仮定に基づき，従来手法における VAE の潜在変数は入力音声の発話内容を表すことが期待できる．しかし，式 (2) において正則化の役割を持つ KL ダイバージェンスの過度な影響により，変換音声の発話内容が消失する傾向にある．この問題は潜在変数の過剰な正則化 [7] として知られており，音声の発話内容が従う複雑な分布が過度に単純化される．また，VAE はその性質上，不特定多数の話者を用いた学習による潜在変数の獲得が可能だが，従来の VAE 音声変換では，学習データに含まれる話者の変換のみに限定されている．

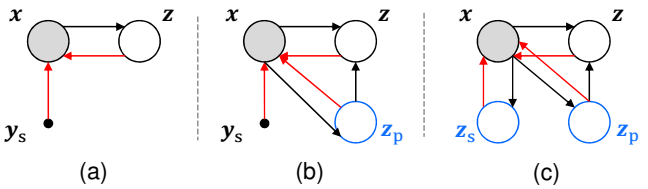

図 2　VAE 音声変換の有向グラフィカルモデル．(a) 話者コードで条件付けされた VAE，(b) 話者コードと音素事後確率で条件付けされた VAE，(c) 話者コードと d-vector で条件付けされた VAE．黒矢印と赤矢印はそれぞれ潜在変数の推論過程と音声パラメータの生成過程を表す．

Fig. 2　Directed graphical models of VAE-based VC models; (a) VAEs conditioned by one-hot speaker codes $\boldsymbol{y}_s$, (b) VAEs conditioned by one-hot speaker codes $\boldsymbol{y}_s$ and PPGs $\boldsymbol{z}_p$, and (c) VAEs conditioned by d-vectors $\boldsymbol{z}_s$ and PPGs $\boldsymbol{z}_p$. Black and red arrows denote inferring latent variables and generating speech parameters $\boldsymbol{x}$, respectively.

3. 提案する VAE 音声変換

本稿では，VAE 音声変換の品質を改善させるための学習法を新たに提案する（3.1 節）．さらに，従来の VAE 音声変換を，学習データに含まれない任意話者対の変換も可能な多対多音声変換へ拡張する（3.2 節）．

3.1 話者コードと音素事後確率で条件付けされた VAE 音声変換

提案手法では，入力音声の発話内容を潜在変数として推定する代わりに，VAE 学習時の補助特徴量として使用する．最も直接的な手法は発話内容の音素系列を用いることだが，本稿では学習済みの音声認識 (Automatic Speech Recognition: ASR) モデル $R(\cdot)$ の出力として得られる音素事後確率 [9] を用いる．$\boldsymbol{z}_p = R(\boldsymbol{x})$ を音声パラメータ $\boldsymbol{x}$ から推定される音素事後確率とすると，式 (2) に示す変分下限は次式で書き換えられる．

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x},\boldsymbol{y}_s,\boldsymbol{z}_p) = &- D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{z}_p)\,||\,p_{\boldsymbol{\theta}}(\boldsymbol{z})\right)\\ &+ \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{z}_p)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{z}_p,\boldsymbol{y}_s)\right]\end{aligned} \tag{3}$$

即ち，音素事後確率 $\boldsymbol{z}_p$ は encoder と decoder の両方に入力され，入力音声の発話内容を保持する役割を持つ．$\boldsymbol{z}_p$ は不特定多数の話者で学習された音声認識モデルから得られるため，発話内容を表す話者非依存な潜在変数として解釈できる．図 2(b) に話者コードと音素事後確率で条件付けされた VAE の有向グラフィカルモデルを示す．

3.2 VAE を用いた多対多音声変換のための話者表現推定法

多対多音声変換での変換先話者が学習データに含まれない場合，当該話者による少数の発話データを用いて話者表現を推定する必要がある．本稿では，未知話者に対する話者表現の推定法として (1) 話者コードを未知話者に適応させる手法，及び (2) 話者認証 (Automatic Speaker Verification: ASV) で有効な d-vector を話者表現の潜在変数として新たに用いる手法を比較する．

3.2.1 Backpropagation を用いた話者コードの適応

DNN を用いた多人数話者のテキスト音声合成 [12] において，話者コードを未知の話者に適応させる手法が提案されている．まず，未知話者の話者コードの初期値を $\hat{y}_s^{(\mathrm{tar})}(k) = 1/S$ として設定する．次に，未知話者の音声パラメータ $\boldsymbol{x}^{(\mathrm{tar})}$ と，VAE により復元された音声パラメータ $\hat{\boldsymbol{x}}^{(\mathrm{tar})}$ の間の Mean Squared Error (MSE) $L_{\mathrm{MSE}}(\boldsymbol{x}^{(\mathrm{tar})},\hat{\boldsymbol{x}}^{(\mathrm{tar})}) = (\boldsymbol{x}^{(\mathrm{tar})}-\hat{\boldsymbol{x}}^{(\mathrm{tar})})^\top(\boldsymbol{x}^{(\mathrm{tar})}-\hat{\boldsymbol{x}}^{(\mathrm{tar})})$ を計算する．ここで，$\hat{\boldsymbol{x}}^{(\mathrm{tar})}$ は $p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(\mathrm{tar})}|\hat{\boldsymbol{z}},\hat{\boldsymbol{y}}_s^{(\mathrm{tar})})$ から抽出され，潜在変数 $\hat{\boldsymbol{z}}$ は $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(\mathrm{tar})})$ から抽出される．最後に，MSE の話者コードでの勾配 $\partial L_{\mathrm{MSE}}/\partial\hat{\boldsymbol{y}}_s^{(\mathrm{tar})}$ を backpropagation により計算し，未知話者の話者コードを $\hat{\boldsymbol{y}}_s^{(\mathrm{tar})} - \eta\partial L_{\mathrm{MSE}}/\partial\hat{\boldsymbol{y}}_s^{(\mathrm{tar})}$ として更新する．ここで，η は更新のステップサイズである．上記の計算を反復することで，最終的な話者コードの推定結果を得る．

3.2.2 d-vector の話者表現としての利用

本稿では，学習済みの話者認証モデル $V(\cdot)$ のボトルネック特徴量として得られる d-vector [11] を用いる VAE 音声変換を新たに提案する．$\boldsymbol{z}_s = V(\boldsymbol{x})$ を音声パラメータ $\boldsymbol{x}$ から抽出される d-vector とすると，式 (3) に示す変分下限は次式で書き換えられる．

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x},\boldsymbol{z}_s,\boldsymbol{z}_p) = &- D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{z}_p)\,||\,p_{\boldsymbol{\theta}}(\boldsymbol{z})\right)\\ &+ \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{z}_p)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{z}_p,\boldsymbol{z}_s)\right]\end{aligned} \tag{4}$$

即ち，従来の VAE 音声変換で用いられていた離散的な話者コードは，連続的な d-vector で置き換えられる．話者認証モデルの目的は話者を特定する特徴量を抽出することであるため，d-vector は話者表現の潜在変数として解釈できる．学習時には，d-vector が decoder にフレーム毎に入力される．変換時には，変換先話者の話者表現が有声区間の d-vector の値の平均として推定され，decoder に入力される．これは，有声・無声の区別をせず，全区間の d-vector の値の平均を用いる話者認証での利用法とは異なる．図 2(c) 及び図 3 にそれぞれ音素事後確率と d-vector で条件付けされた VAE 音声変換の有向グラフィカルモデルと提案手法の概略図を示す．

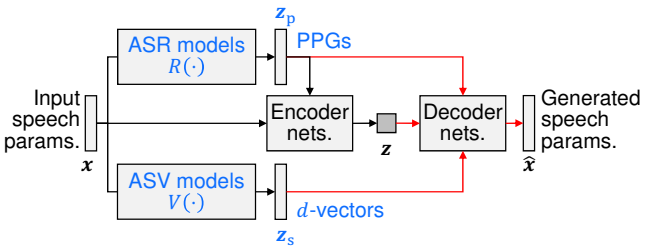

図 3　音素事後確率と d-vector で条件付けされた VAE 音声変換の概略図．ASR と ASV のモデルは，VAE の学習時には更新されない．

Fig. 3　Proposed VAE-based VC conditioned with PPGs and d-vectors. ASR and ASV models are not updated in training for encoder and decoder networks.

3.3 考　　察

提案手法である音素事後確率と d-vector を用いた VAE 音声変換では，ASR と ASV のモデルを構築するための比較的大規模な音声コーパスが必要となる．ここで生じるラベル付けのコストは，条件付き VAE の半教師あり学習 [13] により緩和でき

る．また，end-to-end 音声信号処理 [14,15] の知見を用いた学習も可能である．

4. 実験的評価

4.1 実験条件

実験的評価では，ASR 及び ASV モデル学習用と，音声変換モデル学習・評価用の 2 種類のコーパスを用いる．ASR 及び ASV モデルの学習に用いるコーパスは，日本人話者 260 名 (男性話者 130 名，女性話者 130 名) による計 31 時間の発話データを含む．音声変換モデルの学習及び評価に用いるコーパスは，日本人話者 3 名 (男性話者 2 名，女性話者 1 名) による 425 発話のパラレルデータを含み，400 発話を学習に，25 発話を評価に用いる．ここで，ノンパラレル音声変換の評価を行うために，1 番目から 200 番目までを変換元話者による発話，201 番目から 400 番目までを変換先話者による発話とする．本稿では，男性話者から男性話者への変換，及び，男性話者から女性話者への変換を行う 2 つのモデルを構築する．音声データのサンプリング周波数は 22.05 kHz，フレームシフトは 5 ms である．スペクトル特徴量として STRAIGHT ボコーダ [16] により抽出された 0 次から 39 次のメルケプストラム係数，音源特徴量として F_0，10 帯域の非周期性指標を用いる．学習時には，メルケプストラム係数を次元毎に平均 0，分散 1 に正規化する．変換音声のメルケプストラム係数の 0 次の成分は，入力音声のものをそのまま使用する．F_0 は線形変換し，非周期性指標は入力音声のものを用いる．音声パラメータの生成には，最尤パラメータ生成 [17] を用いる．

本稿では，以下の音声変換モデルを評価する．

FFNN: パラレルデータを用いて学習された Feed-Forward DNN

VAE-SC: 話者コードで条件付けされた VAE [5]

VAE-SC-PPG: 話者コードと音素事後確率で条件付けされた VAE

VAE-DV-PPG: *d*-vector と音素事後確率で条件付けされた VAE

まず，変換元と変換先の 2 名の話者の発話のみで学習及び変換を行う一対一音声変換で上記のモデルを評価する．VAE 音声変換モデルは，完全なノンパラレルデータで学習される．“FFNN” は理想的な条件下でのベースラインとして用いられ，Dynamic Time Warping (DTW) によってアライメントされた完全なパラレルデータで学習される．また，本稿では，多対多音声変換における “VAE-SC-PPG” 及び “VAE-DV-PPG” の性能も評価する．多対多音声変換のための VAE の学習時には，ASR と ASV のモデルの学習に用いた 260 名の話者を含むコーパスを使用し，評価に用いる 2 名の話者による発話は用いない．変換時には，3.2.1 節及び 3.2.2 節で述べた手法を用いて変換先話者の話者表現を推定する．

DNN と VAE のアーキテクチャは，全て Feed-Forward 型とする．ASR モデルは，56 次元の日本語音素事後確率をフレーム毎に予測する．ASR モデルの隠れ層数は 4，隠れ層の素子数は 1024 である．ASV モデルは，261 次元の話者事後確率 (260 話者と無音区間) をフレーム毎に予測する．ASV モデルの隠れ層数は 4，隠れ層の素子数は 256 である．*d*-vector を抽出するボトルネック層の素子数は 16 である．ASR モデルと ASV モデルの隠れ層の活性化関数は，sigmoid 関数である．VAE の encoder の隠れ層数は 2 であり，第 1 層と第 2 層の隠れ素子数はそれぞれ 256，128 である．隠れ層の活性化関数は Rectified Linear Unit (ReLU) [18] である．Decoder のアーキテクチャは，encoder と対称である．潜在変数の次元は 64 である．音声変換の理想的なベースラインとして用いる Feed-Forward DNN の隠れ層数は 4，隠れ層の素子数は 128，隠れ層の活性化関数は ReLU である．最適化アルゴリズムとして，学習率を 0.01 とした AdaGrad [19] を用いる．音声変換モデル学習時の反復回数は 25 とする．

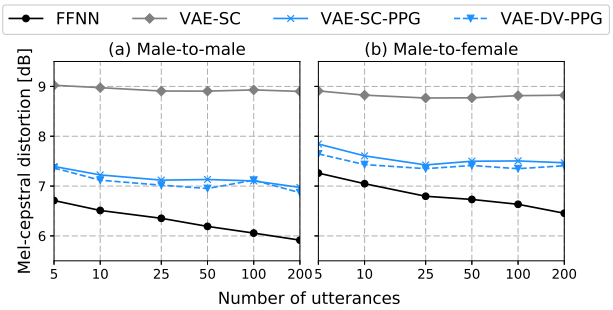

図 4 一対一音声変換におけるメルケプストラム歪み．“FFNN” のモデルのみが，完全なパラレルデータを用いて学習された．

Fig. 4 MCDs of converted speech in one-to-one VC. Only “FFNN” was trained by using fully-parallel speech corpora.

4.2 客観評価

客観評価指標として，自然音声と変換音声のメルケプストラム歪み (Mel-Cepstral Distorsion: MCD) を計算する．自然音声と変換音声のメルケプストラム係数の系列長は，DTW によりアライメントする．本稿では，一対一音声変換における音声変換モデル学習時の発話数と，多対多音声変換における変換先話者の話者表現推定時の発話数を 5, 10, 25, 50, 100, 200 と変化させ，発話数が客観評価指標に与える影響を調査する．

図 4 に一対一音声変換における評価結果を示す．学習時に完全なノンパラレルデータを用いているのにもかかわらず，提案手法である “VAE-SC-PPG” 及び “VAE-DV-PPG” の MCD は，従来の “VAE-SC” と比較して大幅に改善し，完全なパラレルデータを用いて学習された “FFNN” に近づいていることが確認できる．また，話者表現の違いに着目すると，“VAE-DV-PPG” の MCD は “VAE-SC-PPG” と比較してわずかに改善しており，VAE 音声変換において連続的な話者表現である *d*-vector を用いることの有効性を示唆している．

図 5 に多対多音声変換における評価結果を示す．話者表現の推定に用いた発話数に着目すると，“VAE-SC-PPG” の MCD は，発話数を増やすことでわずかに減少する傾向にあることが確認できる．一方で，“VAE-DV-PPG” の MCD は，発話数と話者の性別に依らず，常に “VAE-SC-PPG” よりも低い値とな

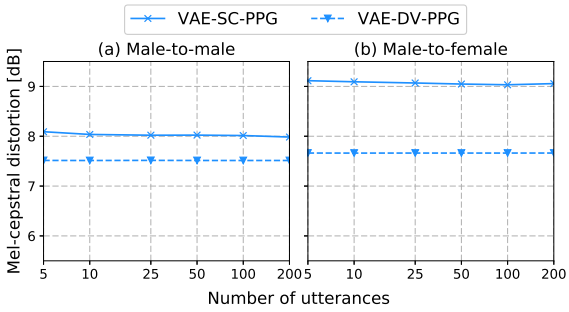

図 5　多対多音声変換におけるメルケプストラム歪み．

Fig. 5　MCDs of converted speech in many-to-many VC.

ることが確認できる．これらの結果は，VAE を用いたノンパラレル多対多音声変換では，離散的な話者コードを適応させる手法よりも，連続的な話者表現である d-vector の利用がより有効であることを示唆している．

4.3　主 観 評 価

提案手法の有効性を検証するために，評価者数を 8 名とした変換音声の自然性に関する Mean Opinion Score (MOS) テスト及び話者類似度に関する Degradation MOS (DMOS) テストを実施する．DMOS テストにおけるリファレンス音声は，変換先話者による発話の分析再合成音声とする．ここでは，“FFNN,” “VAE-SC,” “VAE-SC-PPG (one-to-one),” “VAE-DV-PPG (one-to-one),” “VAE-SC-PPG (many-to-many),” 及び “VAE-DV-PPG (many-to-many)” の 6 手法を同時に評価する．“FFNN” の学習には，変換元と変換先話者による 400 発話のパラレルデータを用いる．“VAE-SC,” “VAE-SC-PPG (one-to-one),” 及び “VAE-DV-PPG (one-to-one)” の学習には，変換元と変換先話者による 200 発話のノンパラレルデータ（計 400 発話）を用いる．多対多音声変換において話者表現の推定に用いる発話数は 100 とする．

図 6 に評価結果を示す．ここで，“FFNN” のみが完全なパラレルデータで学習されており，理想条件下で音声変換モデルを構築した際のベースラインを表している．一対一音声変換における評価結果に着目すると，提案手法である “VAE-SC-PPG (one-to-one)” 及び “VAE-DV-PPG (one-to-one)” は，従来の “VAE-SC” と比較して自然性と話者類似度の両方に関して著しい改善が確認できる．故に，VAE 音声変換において音素事後確率による条件付けを用いることによる有効性が示された．また，多対多音声変換における評価結果に着目すると，変換元と変換先の話者による発話データが音声変換モデル学習時のコーパスに含まれていないにもかかわらず，“VAE-SC-PPG (many-to-many)” 及び “VAE-DV-PPG (many-to-many)” のスコアは，一対一音声変換における提案手法のものと同程度であることが確認できる．また，話者表現の違いに着目すると，d-vector は同姓話者の変換において自然性の改善に有効であることが確認できる．これらの結果より，従来の VAE を用いたノンパラレル音声変換は，効果的に推定された未知話者の話者表現を用いることで多対多音声変換に拡張できることが示された．

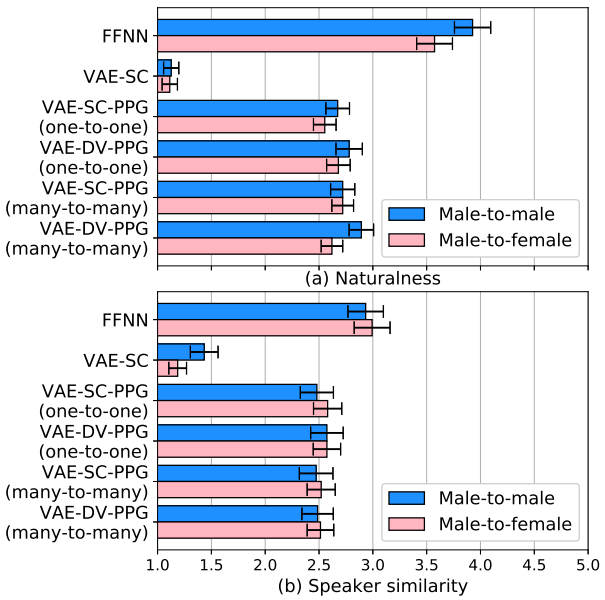

図 6　主観評価結果．(a) 変換音声の自然性に関する MOS スコア，(b) 変換音声の話者類似度に関する DMOS スコア．エラーバーは 95%信頼区間を表す．

Fig. 6　Results of subjective evaluations in terms of (a) naturalness and (b) speaker similarity with 95% confidence intervals.

5.　お わ り に

本稿では，従来の VAE を用いたノンパラレル音声変換の品質を改善させる手法として，学習済みの音声認識モデルの出力として得られる音素事後確率を用いた学習法を提案した．さらに，VAE 音声変換を多対多音声変換に拡張するための効果的な話者表現として，学習済みの話者認証モデルのボトルネック特徴量として得られる d-vector を用いた学習・変換法も新たに提案した．実験的評価結果より，(1) 音素事後確率の導入により，変換音声の自然性及び話者性が劇的に改善すること，及び (2) 話者コードの適応及び d-vector の導入により，任意話者対での変換が可能なノンパラレル多対多音声変換が実現可能となることを示した．今後は，d-vector の次元数や，音声認識・話者認証モデルの性能が変換音声の品質に与える影響を調査する．

謝辞: 本研究は，JSPS 科研費 16H06681 の支援を受けた．

文　　献

[1] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar. 1988.

[2] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, “Voice conversion using artificial neural networks,” in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 3893–3896.

[3] T. Toda, A. W. Black, and K. Tokuda, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[4] D. P. Kingma and M. Welling, “Auto-encoding variational

bayes," *arXiv*, vol. abs/1312.6114, 2013.

[5] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. APSIPA ASC*, Jeju, South Korea, Dec. 2016.

[6] T. Nakashika, T. Takiguchi, and Y. Minami, "Non-parallel training in voice conversion using an adaptive restricted boltzmann machine," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2032–2045, Nov. 2016.

[7] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," *arXiv*, vol. abs/1511.06349, 2016.

[8] N. Dilokthanakul, P. A. M. Mediano, M. Garnelo, M. C. H. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, "Deep unsupervised clustering with Gaussian mixture variational autoencoders," *arXiv*, vol. abs/1611.02648, 2016.

[9] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. ICME*, Seattle, U.S.A., Jul. 2016.

[10] N. Hojo, Y. Ijima, and H. Mizuno, "An investigation of dnn-based speech synthesis using speaker codes," in *Proc. INTERSPEECH*, California, U.S.A., Sep. 2016, pp. 2278–2282.

[11] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 4080–4084.

[12] H.-T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, "Adapting and controlling dnn-based speech synthesis using input codes," in *Proc. ICASSP*, New Orleans, U.S.A., Mar. 2017, pp. 1905–1909.

[13] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. NIPS*, Montreal, Canada, Dec. 2014, pp. 3581–3589.

[14] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," *arXiv*, vol. abs/1701.02720, 2017.

[15] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proc. ICASSP*, Shanghai, China, Mar. 2016, pp. 5115–5119.

[16] H. Kawahara, I. Masuda-Katsuse, and A. D. 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, Apr. 1999.

[17] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, Istanbul, Turkey, June 2000, pp. 1315–1318.

[18] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. AISTATS*, Lauderdale, U.S.A., Apr. 2011, pp. 315–323.

[19] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, Jul. 2011.