Vol.2018-MUS-119 No.28
Vol.2018-SLP-122 No.28
2018/6/16

Variational Autoencoderによる話者ベクトル空間の構築とそれに基づくパラレルデータフリー話者変換

杉山 普[1,a)]　齋藤 大輔[2,b)]　峯松 信明[2]

概要：条件付き Variational Autoencoder (CVAE) を用いて話者情報をベクトル表現へ変換する手法，およびこの話者空間上のベクトルを条件ラベルとする CVAE による声質変換法を提案する．提案手法ではパラレルデータを用いず，声質変換の入出力話者と無関係の多数話者のデータで事前学習された CVAE を声質変換に利用するため，入出力話者の学習データが少量であっても十分な変換精度が期待できる．実験を行った結果，特に入出力話者の学習が少量の場合に，従来の CVAE を用いた声質変換手法と比較して，提案手法が高い変換精度を示した．

1. はじめに

声質変換，特に話者変換とは，ある話者（入力話者）による発話音声をその言語的内容を維持したまま別の話者（出力話者）の音声に変換する技術である．話者変換の応用は音声アプリケーションの出力音声のカスタマイズや 映画のアフレコや吹き替えなど多岐にわたる．

従来の話者変換の多くはまず入力話者と出力話者による同一内容の発話のパラレルデータを用いて入力音声から出力音声への特徴量のマッピングを学習する．学習したマッピングにより入力話者の任意の発話を変換することができる．統計的変換手法としてこれまでベクトル量子化によるコードブックマッピング法や混合ガウス分布モデル（Gaussian Mixture Model; GMM）を用いた手法が提案されてきた [1], [2]．また，近年多くの分野で成果を上げている Artificial Neural Network (ANN) やその応用である Deep Neural Network (DNN) を話者変換におけるマッピングに用いる手法も提案されている [3].

従来手法の多くは学習の際に入力話者と出力話者のパラレルデータが多量に必要となる．しかしこの要件は実用上不都合な場合が多い．例えば何らかの理由で入出力話者とコンタクトが取れず新たな音声の収録ができない場合は学習ができない．また音声アプリケーションの出力の声質をユーザ自身の声質としたい場合，学習データを得るためユーザに多くの文の読み上げ等の労力を強いることになる．このため近年では学習に入出力話者の発話パラレルデータを必要としない手法や少量の学習データでも良い精度を得られる手法が研究されている [4], [5], [6].

[6] で提案されている、Variational Autoencoder (VAE) を用いたパラレルデータフリーな話者変換手法では、まず入出力話者を含む数名の話者の発話データで VAE を学習する．その際デコーダに対して潜在変数だけでなくデータの話者ラベルも与える Conditional VAE (CVAE) として学習する．これによりデコーダの出力する音声の話者性をデコーダに与える話者ラベルによって制御することができる．話者変換の段階では入力話者の音声をエンコードし，得られた潜在変数と出力話者の話者ラベルをデコードすることで所望の変換音声を得る．この手法では CVAE を学習するために話者変換の入出力話者の音声データが多く必要となる．

本研究では CVAE の学習に用いていない話者に対応する話者ラベルをその話者の少量の発話データから推定する手法を提案する．これにより，話者変換の入出力話者と無関係の話者のデータセットで学習された CVAE と，入出力話者の少量の訓練データにより話者変換を行うことができる．

2 章で関連研究を紹介し，3 章で提案手法について述べる．4 章で提案手法を評価する実験について述べ，5 章で本稿をまとめる．

2. 関連研究

2.1 Variational Autoencoder

Variational Autoencoder (VAE) は観測データ x が潜在変数 z をもとに生成されたと仮定し，生成分布 $p(x|z)$ と，

[1] 東京大学大学院情報理工学系研究科
[2] 東京大学大学院工学系研究科
[a)] sugi@tkl.iis.u-tokyo.ac.jp
[b)] dsk_saito@gavo.t.u-tokyo.ac.jp

Vol.2018-MUS-119 No.28
Vol.2018-SLP-122 No.28
2018/6/16

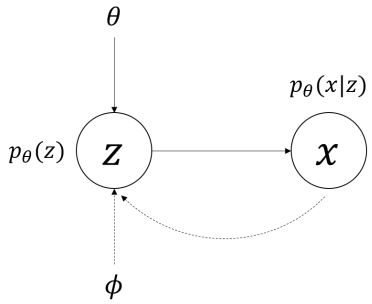


図 1: VAE の生成モデル

事後分布 $p(z|x)$ を DNN により表現する手法である [7].

θ をパラメータとする生成モデル $p_\theta(x) = p_\theta(x|z)p_\theta(z)$ を考える．このモデルの事後分布 $p_\theta(z|x)$ を，ϕ をパラメータとする事後分布モデル $q_\phi(z|x)$ によって近似する (図 1). VAE の学習では以下の式で表される $\log p_\theta(x)$ の変分下限（Evidence Lower Bound; ELBO）を最大にするようなパラメータ θ と ϕ を求める．

$$\log p_\theta(x) \geq \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - D_{KL}\left[q_\phi(z|x)||p_\theta(z)\right] \tag{1}$$

すなわち最小化すべき損失関数 L_{VAE} は以下のようになる．

$$\begin{aligned} &L_{\mathrm{VAE}}(\theta,\phi) \\ &= D_{KL}\left[q_\phi(z|x)||p_\theta(z)\right] - \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right] \end{aligned} \tag{2}$$

本研究では $q_\phi(z|x)$, $p_\theta(x|z)$ はどちらも対角共分散行列を有する正規分布に従うものとし，$p_\theta(z)$ は標準正規分布とする．VAE では $q_\phi(z|x)$ および $p_\theta(x|z)$ をそれぞれエンコーダ，デコーダと呼ばれる DNN により表現する．すなわちエンコーダは x を入力として $q_\phi(z|x)$ の平均 $\mu_z(x)$ と分散 $\Sigma_z(x)$ を出力し，デコーダは z を入力として $p_\theta(x|z)$ の平均 $\mu_x(z)$ と分散 $\Sigma_x(z)$ を出力する (図 2，図 3)．損失関数 L_{VAE} の KL-divergence の項はエンコーダの出力 $\mu_z(x)$, $\Sigma_z(x)$ を用いて以下の閉じた式で書ける．

$$\begin{aligned} &D_{KL}\left[q_\phi(z|x)||p_\theta(z)\right] \\ &= D_{KL}\left[\mathcal{N}(z;\mu_z(x),\Sigma_z(x))||\mathcal{N}(z;0,I)\right] \\ &= \frac{1}{2}\left[\mathrm{tr}\Sigma_z(x) + \mu_z(x)^T\mu_z(x) - k - \log\det\left(\Sigma_z(x)\right)\right] \end{aligned} \tag{3}$$

L_{VAE} の期待値の項は $q_\phi(z|x)$ から J 個の z $(z^{(1)},...,z^{(J)})$ をサンプリングし，それらを用いて以下のように近似する．

$$\mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right] \approx \frac{1}{J}\sum_{j=1}^{J}\log p_\theta(x|z^{(j)}) \tag{4}$$

確率的勾配降下法による学習では $J = 1$ としてよい．$q_\phi(z|x)$ から z をサンプリングするときは，標準正規分布からサンプリングされた ϵ を用いて $z = \mu_z + \sigma_z \circ \epsilon$ ($\circ$ はベクトルの要素ごとの積を表す) とする．これにより学習時にエンコーダへ誤差逆伝播することが可能となる．

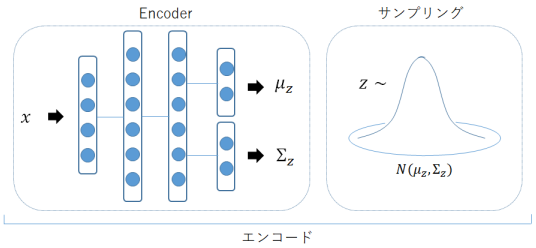


図 2: エンコーダとサンプリングによる推論

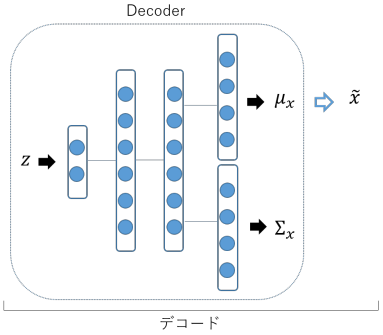


図 3: デコーダによる生成

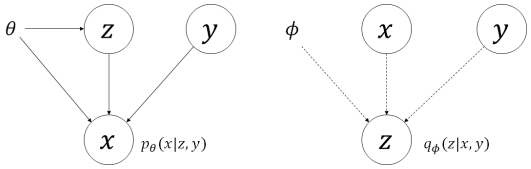


図 4: CVAE の生成モデル

2.2 Conditional Variational Autoencoder

Conditional Variational Autoencoder (CVAE) は，入力データのクラスラベル（人の顔画像であれば性別，音声であれば話者を識別する話者ラベル）が既知である場合に，潜在変数のラベルで条件付けられた事後分布と生成分布を DNN で表現する，VAE の拡張である．図 4 のように観測データ x が，潜在変数 z とラベル y から生成されたとするモデルを考える．データ x の生成確率は以下の式で表される．

$$p_\theta(x) = p_\theta(x|z,y)p_\theta(z)p(y) \tag{5}$$

このモデルにおける z の真の条件付き事後分布 $p_\theta(z|x,y)$ を $q_\phi(z|x,y)$ によって近似する．このとき $p_\theta(x,y)$ の変分下限は，

$$\begin{aligned} &\log p_\theta(x,y) \\ &\geq \mathbb{E}_{z\sim q_\phi(z|x,y)}\left[\log p_\theta(x|z,y)\right] + \log p(y) \\ &\qquad - D_{KL}\left[q_\phi(z|x,y)||p_\theta(z)\right] \\ &= \log p(y) - L_{\mathrm{CVAE}}(\theta,\phi) \end{aligned} \tag{6}$$

となる．ただし，

Vol.2018-MUS-119 No.28
Vol.2018-SLP-122 No.28
2018/6/16

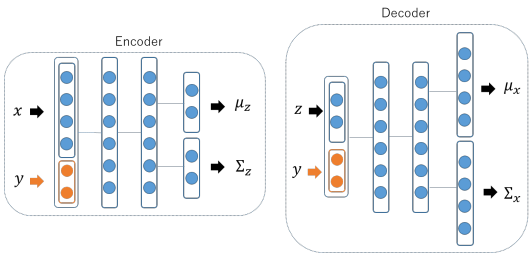


図 5: CVAE のエンコーダおよびデコーダ

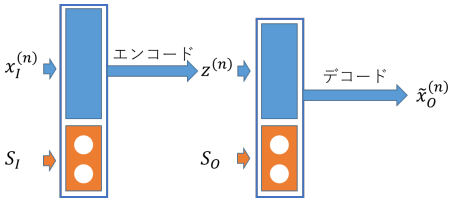


図 6: CVAE を用いた話者変換

$$L_{\mathrm{CVAE}} = D_{KL}\left[q_\phi(z|x,y)||p_\theta(z)\right] - \mathbb{E}_{z\sim q_\phi(z|x,y)}\left[\log p_\theta(x|z,y)\right] \quad (7)$$

である．CVAE の学習では L_{CVAE} を θ と ϕ に関して最小化する．VAE と同様，z の事前分布 $p_\theta(z)$ は標準正規分布，$p_\theta(z|x,y), q_\theta(x|z,y)$ は対角行列を共分散行列とする正規分布を仮定する．ネットワークの構造は VAE のエンコーダとデコーダの入力にクラスラベルを追加したものになる (図 5)．

CVAE では z と y が独立であることを仮定しているため，例えば 0-9 の数字の画像をデータ x，数字のクラス（0-9）をラベル y として学習することで，クラスと筆跡を y と z に分離することができる [8]．音声に適用した場合，話者コードをラベルとして学習することで，潜在変数は言語情報を表現し，ラベルによってデコーダが出力する音声の話者性を制御することが可能であると考えられる [6]．

2.3 CVAE による声質変換

入力話者および出力話者の学習用データをそれぞれ $X_I = \{x_I^{(n)}\}_{n=1}^{N_I}$, $X_O = \{x_O^{(n)}\}_{n=1}^{N_O}$ とする．これを合わせて全学習用データ $X = \{x^{(n)}\}_{n=1}^{N=N_I+N_O}$ とする．また，入出力話者に対して 2 次元の one-hot ベクトル s_I, s_O を話者ラベル（話者ベクトル）として割り当てる．X に対応する話者ラベルの系列を $S = \{s^{(n)}\}_{n=1}^{N}$ とする．話者ラベルは通常 one-hot ベクトルで表現される．入力データを X，入力ラベルを S として CVAE を学習する．こうすることでエンコーダはラベルを考慮してデータの話者情報を取り除き，言語情報を表現する潜在変数を出力することが期待される．また，デコーダは潜在変数と話者ラベルから音声を合成する．

話者変換では学習済みの CVAE を用いて入力話者のデータ x_I と S_I をエンコードし，得られた z と x_O をデコードすることで変換音声 $\tilde{x}_O$ を得る (図 6)．

ここでは入出力話者 2 名のみのデータで CVAE を学習するものとして説明したが，実際には 3 人以上の話者のデータで学習してもよい．この場合学習に用いた任意の話者間で話者変換ができる．

3. 提案手法

3.1 話者ベクトル空間の構築

2.3 節で述べたように話者のベクトル表現として one-hot ベクトルを用いた場合，学習データに含まれない話者に対応するベクトルを考えることができない．本節では話者ベクトル空間内で話者ベクトル s が事前分布 $p(s) = \mathcal{N}(O, I)$ に従って連続的に分布するような話者の埋め込みの方法について説明する．

L 人の話者の音声データからなるデータセット $X = \{x_l^{(n)}\}_{l=1,n=1}^{L,N_l}$ と，対応する言語ラベル $W = \{w_l^{(n)}\}_{l=1,n=1}^{L,N_l}$ が与えられているとする．ただし，N_l は l 番目の話者のデータ数を表す．また言語ラベルは音声の話者に依存しない言語内容を表すものである．X を入力データ，W を入力ラベルとして多人数 CVAE : $\mathrm{CVAE}^{(\mathrm{speaker})}$ を学習する．これにより潜在変数は入力音声の話者性を表現する，話者ベクトルとみなせる．またエンコーダは入力データごとにそれを発話した話者ベクトルの事後分布（平均 $\mu_s(x,w)$，分散 $\Sigma_s(x,w)$ の正規分布）を与えるものと解釈できる．l 番目の話者の全データセット $\{x_l^{(n)}\}_{n=1}^{N_l}$, $W = \{w_l^{(n)}\}_{n=1}^{N_l}$ に対する話者ベクトル s_l は以下のようにエンコーダの出力を用いて最大事後確率推定することができる．

$$\begin{aligned}
\hat{s}_l &= \arg\max_s p(x_l^{(1)}, w_l^{(1)}, ..., x_l^{(N_l)}, w_l^{(N_l)}|s)p(s) \\
&\approx \arg\max_s p(s) \prod_{n=1}^{N_l} p(x_l^{(n)}, w_l^{(n)}|s) \\
&\approx \arg\max_s p(s) \prod_{n=1}^{N_l} \frac{q_\phi(s|x_l^{(n)}, w_l^{(n)})}{p(s)} \\
&= \left[\sum_{n=1}^{N_l} (\Sigma_s(x_l^{(n)}, w_l^{(n)}))^{-1} - (N_l - 1)\boldsymbol{I}\right]^{-1} \\
&\quad \left[\sum_{n=1}^{N_l} \Sigma_s(x_l^{(n)}, w_l^{(n)})^{-1} \mu_s(x_l^{(n)}, w_l^{(n)})\right] \qquad (8)
\end{aligned}$$

このようにして L 人の話者それぞれについて推定された話者ベクトルを，平均 $\boldsymbol{O}$，分散 $\boldsymbol{I}$ となるように正規化したものをこれらの話者の埋め込み表現とする．学習に用いる話者数 L が大きければこの話者ベクトル空間により学習に用いていない話者をもモデル化できることが期待される．

$\mathrm{CVAE}^{(\mathrm{speaker})}$ の学習には音声データセット X に加えて，対応する言語ラベル W が必要である．本研究では L 次元の one-hot 表現の話者ベクトルをラベルとする CVAE : $\mathrm{CVAE}_{\mathrm{OH}}^{(\mathrm{lang})}$ を先に学習し，そのエンコーダが出力する潜在変数 Z を X に対応する言語ラベルとみなして $\mathrm{CVAE}^{(\mathrm{speaker})}$

Vol.2018-MUS-119 No.28
Vol.2018-SLP-122 No.28
2018/6/16

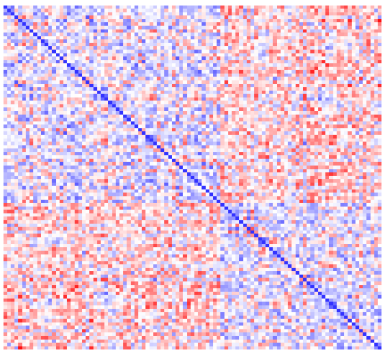


図 7: 101 名の話者に割り当てられた話者コードのコサイン類似度行列．$[-1, 1]$ のスケールを赤 (-1)，白 (0)，青 (1) のグラデーションで表示している．行の左から 58 名が女性，43 名が男性．

の学習に用いる．

実際に VCTK-Corpus[*1]中の 101 名の話者に対して，上記の方法で 12 次元の話者ベクトルを割り当てた．そのコサイン類似度行列を図 7 に示す．割り当てた話者コードが単に話者を識別するだけでなく，性別に起因する音声の特徴を捉えられていることがわかる．

3.2 未知話者の話者ベクトル推定

CVAE の学習データに含まれない話者に対応する話者ベクトルを推定する手法について述べる．

3.1 節の方法により話者ベクトル $\{s_l\}_{l=1}^{L}$ が割り当てられた L 人の学習話者の音声データセット $X = \{x_n\}_{n=1}^{N}$ と，対応する話者ベクトル $S = \{s_{l_n}\}_{n=1}^{N}$ を用いて，$\mathrm{CVAE}_N^{(\mathrm{lang})}$ を学習する．学習用データに存在しない未知話者の音声データ $\dot{X} = \{\dot{x}^{(t)}\}_{t=1}^{T}$ が与えられた時，この話者の話者ベクトル $\dot{s}$ を次のように最大事後確率推定する．

$$\begin{aligned}\dot{s} &= \arg\max_s p(s|\dot{x}^{(1)}, ..., \dot{x}^{(T)}) \\ &= \arg\max_s p(\dot{x}^{(1)}, ..., \dot{x}^{(T)}|s)p(s) \qquad (9) \\ &\approx \arg\max_s p(s)\prod_{t=1}^{T} p(\dot{x}^{(t)}|s) \qquad (10)\end{aligned}$$

ここで，

$$\begin{aligned}p(\dot{x}^{(t)}|s) &= \int_{Z_t} p(\dot{x}^{(t)}|z^{(t)}, s)p(z^{(t)}|s) \\ &= \int_{Z_t} p(\dot{x}^{(t)}|z^{(t)}, s)p(z^{(t)}) \\ &\approx \sum_{j=1}^{J} p_\theta(\dot{x}^{(t)}|z^{(t,j)}, s) \qquad (11)\end{aligned}$$

ここで，$z^{(t)}$ は $\dot{x}$ を生成した潜在変数，Z_t は $z^{(t)}$ の従う分布であり，$z^{(t,j)}$ は Z_t からサンプリングした J 個の潜在変数で j 番目のものを表す．これらをまとめると，

$$\begin{aligned}\dot{s} &\approx \arg\max_s p(s)\prod_{t=1}^{T}\sum_{j=1}^{J} p_\theta(\dot{x}^{(t)}|z^{(t,j)}, s) \\ &= \arg\max_s \mathcal{N}(s; 0, I)\prod_{t=1}^{T}\sum_{j=1}^{J}\mathcal{N}(\dot{x}^{(t)}; \mu_x(z^{(t,j)}, s), \Sigma_x(z^{(t,j)}, s)) \qquad (12)\end{aligned}$$

ここで，$\mu_x(z, s)$, $\Sigma_x(z, s)$ は，潜在変数 z およびラベル s を入力したときにデコーダが出力する正規分布の平均と分散である．$Z_t \approx q_\phi(\dot{x}^{(t)}, s)$，$J = 1$ と近似すると，関数

$$L_{\mathrm{SI}}(s) = -\mathcal{N}(s; 0, I)\prod_{t=1}^{T}\mathcal{N}(\dot{x}^{(t)}; \mu_x(z^{(t)}, s), \Sigma_x(z^{(t)}, s)) \qquad (13)$$

を用いて，

$$\dot{s} \approx \arg\min_s L_{\mathrm{SI}}(s) \qquad (14)$$

と書ける．ただし，式 (13) の $z^{(t)}$ は $q_\phi(\dot{x}^{(t)}, s)$ からサンプリングされた z である．ここで $L_{\mathrm{SI}}(s)$ は s を変数とする DNN によって記述できる．この DNN を s に関する確率的勾配降下法により最適化することで未知話者の話者ベクトル $\dot{s}$ を推定することができる．

3.3 話者変換

3.1 節および 3.2 節で述べた方法により事前に多数の話者の音声データを用いて $\mathrm{CVAE}_N^{(\mathrm{lang})}$ を学習しておく．話者変換の入出力話者の学習データから 3.2 節の方法でそれぞれの話者ベクトルを推定し，推定された話者ベクトルと $\mathrm{CVAE}_N^{(\mathrm{lang})}$ を用いて 2.3 節の方法で話者変換を行うことができる．

4. 評価実験

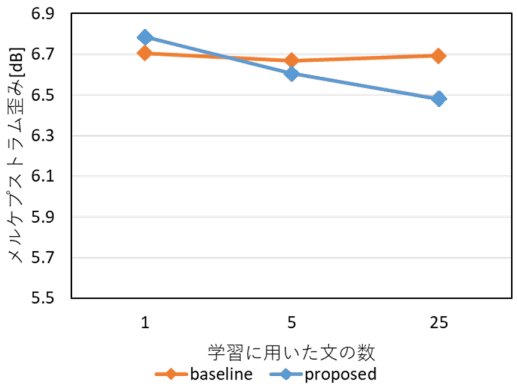
4.1 実験条件

提案手法を評価するため実験を行った．実験系として VCTK-Corpus を用いた．多人数 CVAE ($\mathrm{CVAE}^{(\mathrm{speaker})}$ と $\mathrm{CVAE}_N^{(\mathrm{lang})}$ および $\mathrm{CVAE}_{\mathrm{OH}}^{(\mathrm{lang})}$) の学習用データとして，101 名の話者（各話者 80 文）を用いた．また各話者の 60 文を訓練データ，20 文を開発データとした．話者変換のペアは同性間 2 組 (F2F, M2M) および異性間 1 組 (F2M) とした．これらの話者は多人数 CVAE の学習用データには含まれない．比較のため 2 章で述べた 2 話者 CVAE よる話者変換（従来手法）を実装した．

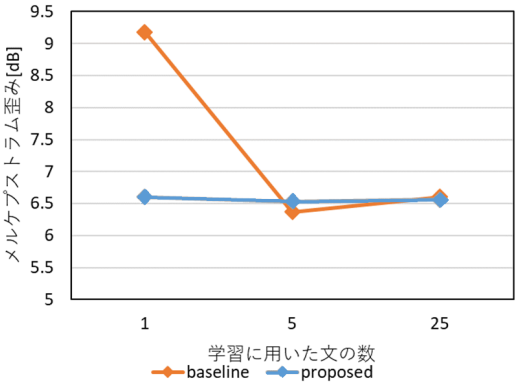
話者変換の入出力話者の学習データは，1 文（2 名で合計で 2 文），5 文（合計で 10 文），25 文（合計で 50 文）の 3 パターンとし，従来手法では 2 話者 CVAE の学習に，提案手法では話者ベクトルの推定にそれぞれ用いた．

WORLD 分析 [9], [10] により音声データの基本周波数，

[*1] http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html



(a) 異性間の話者変換



(b) 同性間の話者変換

図 8: 客観評価結果

非周期成分，スペクトルを抽出した．さらにスペクトルを SPTK[*2] によりパワー項を除いた 24 次元のメルケプストラムに変換し，これを 8 フレーム連結したものを各ネットワークへ入力する特徴量とした．

提案手法の評価として，テストデータ 10 文の変換音声と正解音声とのメルケプストラム歪みの平均による客観評価を行った．また，主観評価として音質に関する AB テストと話者性に関する XAB テストを実施した．主観評価は 6 人で行った．

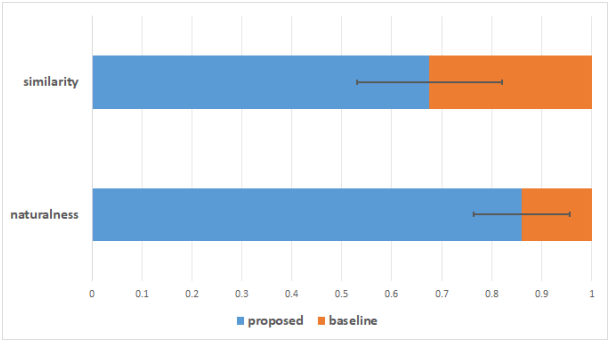
4.2 実験結果

異性間の話者変換の客観評価を図 8(a) に，同性間の話者変換の客観評価を図 8(b) にそれぞれ示す．なお無変換の入力音声と目標音声のメルケプストラム歪みの平均値は異性間で 7.58dB，同性間で 7.99dB であった．

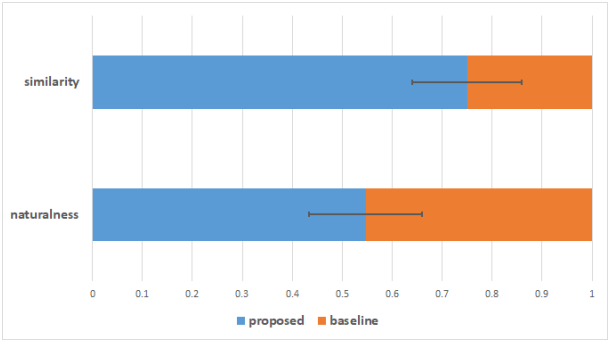
入出力話者それぞれ 5 文で学習した場合の従来手法と提案手法の主観評価による比較を図 9(a) に示す．また，入出力話者それぞれ 25 文で学習した場合の従来手法と提案手法の主観評価による比較を図 9(b) に示す．

主観評価および客観評価から提案手法が従来手法と比較して，特に学習に用いる文が少ない場合に精度よく変換できていると言える．すなわち少量の入出力話者のデータに加えて，入出力話者と無関係の多数の話者で学習した音声の CVAE モデルを話者変換に利用する提案手法の有効性が示唆されている．

[*2] http://sp-tk.sourceforge.net/



(a) 学習データが 5 文ずつの場合



(b) 学習データが 25 文ずつの場合

図 9: 主観評価結果

5. まとめ

本研究ではパラレルデータフリーかつ少量の学習データによる話者変換を目指し，入出力話者とは無関係の多数の話者のデータで学習した CVAE を活用する話者変換の手法を提案した．従来の CVAE による話者変換では CVAE の学習に入出力話者の学習データが多く必要となる．提案手法では事前に多数の話者に対して話者ベクトル空間を構築し疑似的に標準正規分布に従って分布する話者ベクトルを割り当てる．次にこの話者ベクトルをラベルとして多人数 CVAE を学習し，CVAE の学習データに含まれない入出力話者に対して少量の音声データをもとに話者ベクトルを与え，これを利用して話者変換を行う．また実験により従来手法と比較し，提案手法の有効性を示した．

本研究で行った実験では入出力話者の学習データが少量の場合に従来手法と比較して良い変換精度が得られたものの，客観評価指標であるメルケプストラム歪は 6dB を超えており，話者変換の精度として高いとは言えない．このためネットワーク構造やハイパーパラメータを見直すことによる精度の改善が課題となる．また本研究では音声データに対する言語ラベルとして，one-hot 表現の話者ラベルで条件付けした CVAE（$\mathrm{CVAE}_{\mathrm{OH}}^{(\mathrm{lang})}$）で得られる潜在変数を

用いたが，他にも音素事後確率や特定の話者の声質に変換した音声を言語ラベルとして用いることが精度改善の策として考えられる．

さらに，CVAE を用いた話者変換は CVAE の潜在変数が言語情報を表現すること，かつそれが話者ラベルに非依存（例えば/i/を表現する z は話者ラベルの値に依らず/i/を表現する）であることを前提としている．しかし，この性質は CVAE の学習制約から自明に導かれるものではない．提案手法も従来手法と同様，最終的な変換精度はこの性質がどれほど強く成り立っているかに左右されるため，提案手法による話者変換のさらなる精度向上には，CVAE にこの性質を持たせるような明示的な学習制約を検討することが必要となる．

参考文献

[1] Stylianou, Y., Cappé, O. and Moulines, E.: Continuous probabilistic transform for voice conversion, *IEEE Transactions on speech and audio processing*, Vol. 6, No. 2, pp. 131–142 (1998).
[2] Toda, T., Black, A. W. and Tokuda, K.: 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 (2007).
[3] Desai, S., Raghavendra, E. V., Yegnanarayana, B., Black, A. W. and Prahallad, K.: Voice conversion using artificial neural networks, *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, IEEE, pp. 3893–3896 (2009).
[4] 戸田智基，大谷大和，鹿野清宏：固有声に基づく声質変換法，電子情報通信学会技術研究報告. SP, 音声， Vol. 106, No. 221, pp. 25–30 (2006).
[5] Sun, L., Li, K., Wang, H., Kang, S. and Meng, H.: Phonetic posteriorgrams for many-to-one voice conversion without parallel data training, *Multimedia and Expo (ICME), 2016 IEEE International Conference on*, IEEE, pp. 1–6 (2016).
[6] Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y. and Wang, H.-M.: Voice conversion from non-parallel corpora using variational auto-encoder, *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*, IEEE, pp. 1–6 (2016).
[7] Kingma, D. P. and Welling, M.: Auto-encoding variational bayes, *arXiv preprint arXiv:1312.6114* (2013).
[8] Kingma, D. P., Mohamed, S., Rezende, D. J. and Welling, M.: Semi-supervised learning with deep generative models, *Advances in Neural Information Processing Systems*, pp. 3581–3589 (2014).
[9] Morise, M., Yokomori, F. and Ozawa, K.: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, *IEICE TRANSACTIONS on Information and Systems*, Vol. E99-D, No. 7, pp. 1877–1884 (2016).
[10] Morise, M.: D4C, a band-aperiodicity estimator for high-quality speech synthesis, *Speech Communication*, Vol. 84, pp. 57–65 (2016).