音素事後確率と d-vector を用いたノンパラレル多対多 VAE 音声変換における学習データ量と d-vector 次元数に関する評価 *

☆中村 泰貴（東京大学），齋藤 佑樹（東大院・情報理工），
西田 京介，井島 勇祐（NTT），高道 慎之介（東大院・情報理工）

1 はじめに

Variational AutoEncoder (VAE) [1] を用いたノンパラレル音声変換 [2] は，その学習の容易さから，広く研究され始めている．これまでに我々は，音声認識 (Automatic Speech Recognition: ASR) と話者認証 (Automatic Speaker Verification: ASV) の知見を利用したノンパラレル多対多 VAE 音声変換 [3] を提案し，その有効性を確認している．この手法では，ASR モデルの推論結果から得られる音素事後確率 (Phonetic PosteriorGram: PPG) [4] と，ASV モデルの中間表現として得られる d-vector [5] を利用し，不特定多数の話者の音声パラメータを高品質にモデル化・変換可能な VAE 音声変換モデルを学習する．ノンパラレル多対多 VAE 音声変換では，ASR モデルおよび ASV モデルの事前学習と，VAE 音声変換モデルの学習に多数の話者を含む大規模なコーパスを用いる．また，話者空間を表現する d-vector の次元数も学習前に決定する必要がある．本稿では，これらの話者数と次元数が最終的な変換音声の品質に与える影響を主観評価により調査し，(1) 低次元の d-vector 表現が変換音声品質を改善させ，(2) 話者数の増加が変換音声品質を改善させることを示す．

2 ノンパラレル多対多 VAE 音声変換

ノンパラレル多対多 VAE 音声変換 [3] では，音声の発話内容の潜在変数として PPG を，話者の潜在変数として d-vector を導入した VAE 音声変換モデルを学習する．VAE の encoder $p_\theta(\cdot)$ と decoder $q_\phi(\cdot)$ のパラメータ θ と ϕ の推定に用いる変分下限は，次式で書き表される．

$$\begin{aligned}\mathcal{L}(\theta,\boldsymbol{\phi};\boldsymbol{x},\boldsymbol{z}_\mathrm{s},\boldsymbol{z}_\mathrm{p}) = &-D_\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{z}_\mathrm{p})||p_\theta(\boldsymbol{z})) \\ &+\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{z}_\mathrm{p})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{z}_\mathrm{p},\boldsymbol{z}_\mathrm{s})]\end{aligned} \quad (1)$$

ここで，$\boldsymbol{z}_\mathrm{p} = R(\boldsymbol{x})$ と $\boldsymbol{z}_\mathrm{s} = V(\boldsymbol{x})$ はそれぞれ学習済みの ASR モデル $R(\cdot)$ と ASV モデル $V(\cdot)$ を用いて音声パラメータ $\boldsymbol{x}$ から推定される PPG と d-vector である．$\boldsymbol{z}$ は VAE の潜在変数である．Figure 1 にノンパラレル多対多 VAE 音声変換モデルの概略図を示す．ASR モデルと ASV モデルの事前学習に用いる話者数が多いほど，(1) 入力音声の発話内容を保持する PPG が話者非依存な特徴量となり，(2) d-vector が多様な話者を表現可能となることが期待できる．一方で，d-vector の次元数を増加させることで ASV モデルの性能は改善すると予想されるが，話者空間における話者の分布がより疎になることも考えられ，必ずしも変換音声の品質改善に結びつくとは限らない．

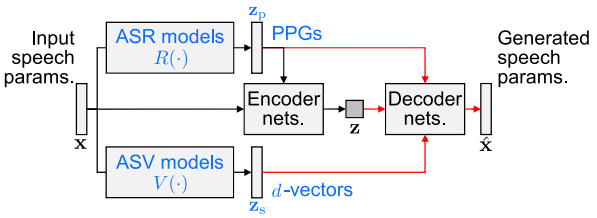

Fig. 1 ノンパラレル多対多 VAE 音声変換モデルの概略図．事前学習済みの ASR と ASV のモデルは，VAE の学習時には更新されない．

3 実験的評価

3.1 実験条件

実験的評価では，(1) ASR モデル，ASV モデル，VAE 音声変換モデル学習用と，(2) VAE 音声変換モデル評価用の 2 種類のコーパスを用いる．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．本稿では，学習時の話者数を (1) 男女 25 名ずつ (“**50spk**”)，(2) 男女 65 名ずつ (“**130spk**”)，そして (3) 男女 130 名ずつ (“**260spk**”) の 3 通りに設定する．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，25 発話を評価に用いる．変換先話者の d-vector の推定には，評価データ以外の 200 発話を用いる．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 [6] ボコーダにより抽出された 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 F0，10 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．F0 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 [7] を用いる．

ASR，ASV，そして VAE 音声変換モデルのニューラルネットワーク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 PPG 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 1,024 である．ASV モデルは，学習に用いた話者と無声区間（話者数 +1 次元の話者事後確率）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．本稿では，d-vector の次元数（即ち，ASV モデルのボトルネック層の素子数）を (1) 8 次元 (“**d8**”)，(2) 16 次元 (“**d16**”)，そして (3) 32 次元 (“**d32**”) の 3 通りに設定する．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

---

* Evaluation of VAE-based non-parallel and many-to-many voice conversion conditioned by phonetic posteriorgrams and d-vectors in terms of training data and dimensionality of d-vectors by NAKAMURA, Taiki, SAITO, Yuki (The University of Tokyo), NISHIDA, Kyosuke, IJIMA, Yusuke (NTT), and TAKAMICHI, Shinnosuke (The University of Tokyo).

Table 1　話者数固定のもとで *d*-vector の 次元数を変えた場合の変換音声の自然性に関する評価結果

spk		d8 vs. d16	d16 vs. d32	d8 vs. d32
50	m2m	**0.548** - 0.452	**0.516** - 0.484	**0.596** - 0.404
	m2f	**0.504** - 0.496	**0.536** - 0.464	**0.536** - 0.464
130	m2m	0.472 - **0.528**	**0.580** - 0.420	**0.568** - 0.432
	m2f	**0.504** - 0.496	**0.584** - 0.416	**0.548** - 0.452
260	m2m	**0.516** - 0.484	**0.556** - 0.444	**0.568** - 0.432
	m2f	**0.564** - 0.436	0.492 - **0.508**	**0.640** - 0.360

Table 2　話者数固定のもとで *d*-vector 次元数を変えた場合の変換音声の話者類似性に関する評価結果

spk		d8 vs. d16	d16 vs. d32	d8 vs. d32
50	m2m	**0.532** - 0.468	**0.532** - 0.468	**0.508** - 0.492
	m2f	0.476 - **0.524**	**0.516** - 0.484	**0.568** - 0.432
130	m2m	0.448 - **0.552**	**0.516** - 0.484	**0.552** - 0.448
	m2f	0.456 - **0.544**	**0.572** - 0.428	**0.580** - 0.420
260	m2m	**0.552** - 0.448	**0.544** - 0.456	0.484 - **0.516**
	m2f	**0.572** - 0.428	0.496 - **0.504**	**0.564** - 0.436

はそれぞれ 256，128 である．隠れ層の活性化関数は Rectified Linear Unit (ReLU) [8] である．Decoder のアーキテクチャは，encoder と対称である．VAE の潜在変数の次元は 64 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 [9] を用いる．音声変換モデル学習時の反復回数は 25 とする．

3.2　主観評価

本稿では，前節で述べた話者数 {50spk, 130spk, 260spk} の計 3 通り，*d*-vector の次元数 {d8, d16, d32} の計 3 通り，{m2m（男性話者から男性話者への変換），m2f（男性話者から女性話者への変換）} の計 2 通り，合計 18 通りの条件のもとで生成された変換音声の品質を評価する．これらの条件のもと，我々のクラウドソーシングによる評価システムを用いて，変換音声の自然性に関するプリファレンス AB テスト及び話者類似性に関するプリファレンス XAB テストを実施する．各評価における受聴者数は 25 人であり，1 人あたり 10 サンプルの音声を評価する．プリファレンス XAB 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とする．総当たりでの評価の実施は困難であるため，まず，話者数を固定させたもとで *d*-vector の次元数に関する評価を行う．その後，各話者数において最適な *d*-vector の次元数を用いた場合の 3 通りの組み合わせで評価を行う．

3.2.1　*d*-vector の次元数に関する評価

Table 1 および Table 2 にそれぞれ話者数を固定させたもとで *d*-vector の次元数を変えた場合の変換音声の自然性に関する主観評価結果および話者類似性に関する主観評価結果を示す． まず，自然性に関する評価結果 (Table 1) より，いくつかの例外はあるものの，全体として *d*-vector の 次元数を増加させることで変換音声の自然性が劣化する傾向があることが確認できる．自然性ほど顕著ではないが，話者類似性に関する評価結果 (Table 2) においても同様の傾向が確認できる．以上より，話者表現である *d*-vector の次元数は，低次元である方が好ましいことが示された．

Table 3　各話者数に対して最適な *d*-vector の 次元数を用いた場合の変換音声の自然性に関する評価結果．Table 1 の結果から，全ての話者数に対して *d*-vector の次元数を 8 とした

	50spk vs. 130spk	130spk vs. 260spk	50spk vs. 260spk
m2m	0.496 - **0.504**	0.472 - **0.528**	0.400 - **0.600**
m2f	0.488 - **0.512**	0.476 - **0.524**	0.380 - **0.620**

Table 4　各話者数に対して最適な *d*-vector の 次元数を用いた場合の変換音声の話者類似性に関する評価結果．Table 2 の結果から，*d*-vector の次元数を 130spk に対して 16，それ以外に対して 8 とした

	50spk vs. 130spk	130spk vs. 260spk	50spk vs. 260spk
m2m	0.448 - **0.552**	0.448 - **0.552**	0.436 - **0.564**
m2f	0.456 - **0.544**	0.472 - **0.528**	0.392 - **0.608**

3.2.2　*d*-vector 次元数および学習話者数に関する評価

Table 3 および Table 4 にそれぞれ話者数に対して最適な *d*-vector の次元数を用いた場合の変換音声の自然性に関する主観評価結果および話者類似性に関する主観評価結果を示す．これらの表より，全ての場合において，話者数を増加させることで変換音声の自然性および話者類似性が改善していることが確認できる．以上より，ノンパラレル多対多 VAE 音声変換において，学習に用いる話者数の増加が変換音声の品質を改善させることが示された．

4　おわりに

本稿では，ノンパラレル多対多 VAE 音声変換における話者数および *d*-vector の次元数が変換音声品質に与える影響を実験的に評価し，評価結果より，(1) 低次元の *d*-vector 表現が変換音声品質を改善させ，(2) 話者数の増加が変換音声品質を改善させることを示した．今後は，変換話者対の影響についてさらに調査する．

参考文献

[1] Kingma et al., *arXiv:1312.6114,* 2013.

[2] Hsu et al., *Proc. APSIPA ASC,* pp. 1–6, 2016.

[3] Saito et al., *Proc. ICASSP*, pp. 5274–5278, 2018.

[4] Sun et al., *Prpc. ICME,* pp. 1–6, 2016.

[5] Variani et al., *Proc. ICASSP,* pp. 4080–4084, 2014.

[6] Kawahara et al., *Speech Communication,* vol. 27, no. 3–4, pp. 187–207, 1999.

[7] Tokuda et al., *Proc. ICASSP,* pp. 1315–1318, 2000.

[8] Glorot et al., *Proc. AISTATS,* pp. 315–323, 2011.

[9] Duchi et al., *The Journal of Machine Learning Research,* vol. 12, pp. 2121–2159, 2011.