Non-parallel and many-to-many voice conversion using variational autoencoders integrating speech recognition and speaker verification

Yuki Saito1,*, Taiki Nakamura2,†, Yusuke Ijima3,‡, Kyosuke Nishida3,§ and Shinnosuke Takamichi1,†

1Graduate School of Information Science and Technology, The University of Tokyo, 7–3–1 Hongo, Bunkyo-ku, Tokyo, 133–8656 Japan
2Faculty of Engineering, The University of Tokyo, 7–3–1 Hongo, Bunkyo-ku, Tokyo, 133–8656 Japan
3NTT Media Intelligence Laboratories, NTT Corporation, 1–1 Hikarinooka, Yokosuka, 239–0847 Japan

(Received 9 December 2019, Accepted for publication 8 July 2020)

Abstract: We propose non-parallel and many-to-many voice conversion (VC) using variational autoencoders (VAEs) that constructs VC models for converting arbitrary speakers’ characteristics into those of other arbitrary speakers without parallel speech corpora for training the models. Although VAEs conditioned by one-hot coded speaker codes can achieve non-parallel VC, the phonetic contents of the converted speech tend to vanish, resulting in degraded speech quality. Another issue is that they cannot deal with unseen speakers not included in training corpora. To overcome these issues, we incorporate deep-neural-network-based automatic speech recognition (ASR) and automatic speaker verification (ASV) into the VAE-based VC. Since phonetic contents are given as phonetic posteriorgrams predicted from the ASR models, the proposed VC can overcome the quality degradation. Our VC utilizes $d$-vectors extracted from the ASV models as continuous speaker representations that can deal with unseen speakers. Experimental results demonstrate that our VC outperforms the conventional VAE-based VC in terms of mel-cepstral distortion and converted speech quality. We also investigate the effects of hyperparameters in our VC and reveal that 1) a large $d$-vector dimensionality that gives the better ASV performance does not necessarily improve converted speech quality, and 2) a large number of pre-stored speakers improves the quality.

Keywords: Non-parallel and many-to-many voice conversion, Variational autoencoders, Phonetic posteriorgrams, $d$-vectors

1. INTRODUCTION

Voice conversion (VC) is a technique to transform the characteristics of source speech into those of target speech while keeping its linguistic information, which can be used for various applications such as language education [1], entertainment [2], and speaking aids [3]. This paper focuses on statistical VC [4] that constructs VC models using speech corpora for the training and enables us to control the characteristics of the converted speech flexibly.

In parallel VC using deep neural networks (DNNs) [5], DNN-based VC models learn a frame-wise mapping function between source and target speech parameters using parallel speech corpora that include the same utterances recorded by the source and target speakers. The DNN-based VC not only can achieve better converted speech quality than conventional Gaussian mixture model (GMM)-based one [6] but also can offer a way to incorporate other DNN-based speech processing techniques such as automatic speech recognition (ASR) [7] and automatic speaker verification (ASV) [8] into the VC framework (e.g., incorporating DNN-based ASV into VC training [9]). However, recording parallel speech corpora is often difficult and tiresome in practice.

One way to overcome the difficulty is using non-parallel VC, which does not require any parallel speech corpora to construct VC models. So far, deep generative models such as variational autoencoders (VAEs) [10,11] and generative adversarial networks [12] using cycle-consistency (a.k.a., CycleGAN) [13–15] have been applied
to non-parallel VC. We focus on VAE-based VC because its training criterion is based on the theoretically grounded variational Bayesian inference, rather than the minimax optimization that is difficult to solve stably. In the conventional VAE-based VC [11], speaker-independent encoder networks infer latent variables of input speech parameters, and speaker-dependent decoder networks conditioned by one-hot speaker codes [16] generate the parameters from the latent variables. Thus, the speaker-independent latent variables can be expected to represent phonetic contents of speech and VC can be simply done by modifying the speaker codes. However, the converted speech quality of VAE-based VC is lower than that of DNN-based parallel VC. One of the reasons for this is an over-regularization effect observed in the latent variables [17], which makes the distribution of the latent variables too simplistic. Using more complex prior distribution of the latent variables such as GMMs [18] can alleviate the effect; however, determining the number of clusters of the GMM prior is likely to be difficult because variation in the phonetic contents is typically large. Moreover, although we can train the VAEs by using various speech corpora, they cannot deal with unseen speakers not included in training data because their speaker codes are not defined.

To deal with the issues in the conventional VAE-based VC, we propose non-parallel and many-to-many VC using VAEs that can convert arbitrary speakers’ characteristics into those of other arbitrary speakers with VC models trained with non-parallel speech corpora. As speech corpora for constructing DNN-based ASR and ASV models can also be used for training VC models based on VAEs, we integrate the three models for achieving non-parallel and many-to-many VC. We utilize phonetic posteriorgrams (PPGs) [19] predicted from the ASR models for dealing with the quality degradation issue. We also use d-vectors [8] extracted from a bottleneck layer of the ASV models for replacing the conventional discrete speaker representations with continuous ones that can identify the input speaker, even if his/her speech data was not used for training the ASV models. Experimental results demonstrate that PPGs successfully improve both naturalness and speaker similarity of the converted speech obtained with VAE-based VC. They also show that the use of d-vectors is more effective for reducing mel-cepstral distortion of the converted speech in the proposed VC than the use of conventional speaker code adaptation. We also investigate the effects of the number of the pre-stored speakers and the dimensionality of the d-vectors in the proposed VC. The investigation results reveal that 1) a large d-vector dimensionality that enhances the performance of the ASV models does not always improve converted speech quality, and 2) a large number of pre-stored speakers and tends to improve converted speech quality. Note that this paper is partially based on an international conference paper [20] written by the authors. The contribution of this paper is that we provide a new analysis of the proposed VC and report enhanced experiments carried out under various conditions.

This paper is organized as follows. Section 2 briefly reviews the conventional VAE-based non-parallel VC. Section 3 describes the VAE-based non-parallel and many-to-many VC we propose. Section 4 presents experimental evaluations. Section 5 concludes the paper with a summary and a mention of future work.

2. CONVENTIONAL VAE-BASED NON-PARALLEL VC

2.1. Non-parallel VC Using VAEs Conditioned by Speaker Codes

The VAEs used for non-parallel VC are probabilistic generative models that generate speech parameters \( x \) from their latent variables \( z \) and speaker representations \( y_s \). In the approach reported by Hsu et al. [11], one-hot speaker codes [16] are adopted to the speaker representations, which use 1-of-\( S \) representation for identifying one of the pre-stored \( S \)-speakers. The speaker codes for the \( i \)th speaker \( y_s^{(i)} \) are defined as:

\[
y_s^{(i)}(k) = \begin{cases} 1 & \text{if } k = i \\ 0 & \text{otherwise} \end{cases} \quad (1 \leq k \leq S).
\]

Under the assumption that \( z \) is independent of \( y_s \), model parameters of the VAEs, \( \theta \), are estimated by maximizing the marginal likelihood of speech parameters conditioned by speaker codes defined as:

\[
p_\theta(x|y_s) = \int p_\theta(x|z, y_s)p_\theta(z)dz,
\]

where \( p_\theta(z) \) is a prior of the latent variables. Since the integral in Eq. (2) is intractable, we introduce two neural networks; speaker-independent encoder networks \( q_\phi(z|x) \) that approximate the true posterior of the latent variables \( p_\theta(z|x) \), and speaker-dependent decoder networks that approximate the true posterior of the speech parameters of a specific speaker \( p_\theta(x|z, y_s) \). \( \phi \) and \( \theta \) are respectively model parameter sets of the encoder and decoder networks. In training the VAEs, we maximize the variational lower bound of the log likelihood defined as:

\[
\mathcal{L}(\theta, \phi; x, y_s) = -D_{KL}(q_\phi(z|x) \| p_\theta(z)) + \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z, y_s)],
\]

where \( D_{KL}(\cdot \| \cdot) \) denotes the Kullback-Leibler (KL) divergence between two distributions. We assume that both the encoder and decoder networks represent the diagonal Gaussian distributions of which the mean and covariance are estimated by the networks. The isotropic Gaussian distribution \( \mathcal{N}(z; 0, I) \) is typically adopted to the prior \( p_\theta(z) \) for obtaining the closed form of the KL term in Eq. (3).
The reparameterization trick [10] is used for the backpropagation algorithm during the training. Figure 1(a) illustrates the directed graphical model of the VAEs.

Note that parallel speech corpora are not required because the VAEs are trained in the same manner as in the auto-encoding process. After the training, VC can be performed by feeding speaker codes of the target speaker into the decoder. For instance, when we convert source speech parameters into those of the \( j \)-th speaker, the speaker codes \( y^{(j)} \) are fed into the decoder frame-by-frame.

2.2. Problems

Since we assume that the latent variables are independent of the speaker codes, they can be expected to represent phonetic contents of speech. However, in the conventional VAE-based non-parallel VC [11], the phonetic contents tend to vanish because of the excessive effects of the prior used in the KL term of Eq. (3) and the converted speech quality is significantly degraded. This issue is known as an over-regularization of the latent variables [17], which makes the latent variables too simplified to represent the underlying structure of the phonetic contents. This effect can be alleviated by using more complex prior distribution such as GMMs [18]; however, training VAEs with the GMM prior is likely to be more difficult than training those with the isotropic Gaussian distribution. In addition to the quality degradation, using the discrete speaker codes as the speaker representations limits this framework to VC that can only deal with pre-stored speakers included in training data, although the VAEs have the potential to extract speaker-independent latent variables from many and unspecified speakers.

3. PROPOSED VAE-BASED NON-PARALLEL AND MANY-TO-MANY VC

Here, we describe the VAE-based non-parallel and many-to-many VC we propose, which integrates DNN-based ASR and ASV models for converting arbitrary speakers’ characteristics into those of other arbitrary speakers.

3.1. VAE-based Non-parallel VC Integrating ASR

To alleviate quality degradation of converted speech due to the over-regularization effect, we directly utilize the phonetic contents of input speech for training VAEs. We adopt PPGs [19] predicted by pretrained ASR models because they can be regarded as speaker-independent latent variables of the phonetic contents. Let \( z_p = R(x) \) be PPGs predicted from speech parameters \( x \) through the ASR models \( R(\cdot) \). The objective function Eq. (3) is rewritten as:

\[
\mathcal{L}(\theta, \phi; x, y, z_p) = -D_{KL}(q_{\phi}(z|x, z_p) \parallel p_{\theta}(z)) + \mathbb{E}_{q_{\phi}(z|x, z_p)}[\log p_{\theta}(x|z, z_p, y)].
\]

Here, the PPGs \( z_p \) are fed into both the encoder and decoder networks, which keeps the phonetic contents of source speech unchanged in the training and conversion stages. Figure 1(b) shows the directed graphical model of the proposed VAE-based VC using PPGs and speaker codes.

3.2. VAE-based Non-parallel and Many-to-many VC Integrating ASR and ASV

We extend the conventional VAE-based one-to-one VC to many-to-many VC that can deal with unseen speakers not included in training corpora. To achieve many-to-many VC, speaker representations fed into the decoder networks must define the identity of the unseen speakers. Although we can use conventional speaker code adaptation [16,21] for the VAE-based VC, we adopt \( d \)-vectors [8] extracted from a bottleneck layer of pretrained ASV models to the speaker representation. Because the ASV models are trained to extract features that can identify a specific speaker, the \( d \)-vectors can be regarded as latent variables of the speaker representations. Let \( z_s = V(x) \) be \( d \)-vectors extracted from speech parameters \( x \) through the ASV models \( V(\cdot) \). The objective function Eq. (4) is rewritten as:

\[
\mathcal{L}(\theta, \phi; x, z_s, z_p) = -D_{KL}(q_{\phi}(z|x, z_p) \parallel p_{\theta}(z)) + \mathbb{E}_{q_{\phi}(z|x, z_p)}[\log p_{\theta}(x|z, z_p, z_s)].
\]

Here, the conventional discrete speaker codes are replaced with continuous \( d \)-vectors. In the training stage, the \( d \)-vectors are fed into the decoder networks frame-by-frame in the same manner as the speaker codes. In the conversion stage, the speaker representations for target speakers are estimated as a \( d \)-vector averaged over all of those in voiced regions since acoustic features in unvoiced regions are less affected by speaker individuality than those in voiced regions [22,23]. Note that this differs from the traditional use of \( d \)-vectors for speaker verification that represents speaker identity as a \( d \)-vector averaged over all of those in all (i.e., voiced and unvoiced) regions.
empirically confirm the effect of the two hyperparameters. In Sect. 4.4, we proposed VC for achieving multilingual VC. Furthermore, multilingual adaptation of ASR models [29] can be introduced to the proposed VC. Moreover, techniques for end-to-end speech processing [25,26] and dual learning of speech synthesis and speech recognition [27,28] can be applied to the proposed VC. Furthermore, VAEs using speaker codes [11] can be introduced to the proposed VC for achieving multilingual VC.

The number of pre-stored speakers and the dimensionality of $d$-vectors are hyperparameters of the proposed VC. For achieving high-quality many-to-many VC, the number of speakers should be large enough to obtain better latent variables of phonetic contents and speaker representations. The dimensionality of $d$-vectors also affects the converted speech quality since it can be regarded as the number of basis in the continuous speaker space. In Sect. 4.4, we empirically confirm the effect of the two hyperparameters.

4. EXPERIMENTAL EVALUATION

4.1. Experimental Conditions

We used two speech corpora for training and evaluating the VAE-based VC: a parallel speech corpus including many speakers and a non-parallel speech corpus including many speakers. The first corpus included six Japanese speakers (three males and three females) who uttered 425 fully-parallel sentences. The second corpus included 260 Japanese speakers (130 males and 130 females) who uttered about 100 non-parallel sentences. The sampling rate of the corpora was 22.05 kHz.

We first evaluated the proposed VAE-based VC in a one-to-one VC setting to investigate the effectiveness of using the ASR models for improving converted speech quality. We constructed VC models for every pair of source/target speakers taken from the first corpus. Since we used the three male and three female speakers for this evaluation, we finally had 12 inter-gender cases: six male-to-male (m2m) and six female-to-female (f2f), and 18 inter-gender cases: nine male-to-female (m2f) and nine female-to-male (f2m). For making a non-parallel setting, we assumed the 1st-through-200th utterances were recorded by the source speaker and the 201st-through-400th utterances were recorded by the target speaker. The parallel 401st-through-425th utterances were used for the evaluation.

After evaluating one-to-one VC, we evaluated our VC in a many-to-many VC setting. Here, we pretrained the ASR and ASV models by using the second corpus. After the pretraining, we trained VAEs that can convert arbitrary speakers’ characteristics into those of other arbitrary speakers by using the second corpus. We evaluated the VAEs by using the first corpus. Note that the six speakers (i.e., the three males and the three females) used for this evaluation were not included in the training data, and we used some utterances of the speakers for estimating the speaker representations.

In the evaluation, we compared the performances of the following four VC models.

- **FFNN**; Feed-Forward DNNs
- **VAE-SC**: VAEs using speaker codes [11]
- **VAE-SC-PPG**: VAEs using speaker codes and PPGs
- **VAE-DV-PPG**: VAEs using $d$-vectors and PPGs

In the one-to-one VC evaluation, the VAEs were trained with a completely non-parallel speech corpus, while the DNNs used in “FFNN” were trained with a fully-parallel speech corpus aligned by using the dynamic time warping (DTW) algorithm, which were referred to the ideal baseline of the VC models. We also evaluated the proposed “VAE-SC-PPG” and “VAE-DV-PPG” in the many-to-many VC setting. In the conversion stage of “VAE-SC-PPG,” speaker representations for the target speaker were estimated by using the speaker code adaptation method [21], which performs backpropagation algorithm to find a speaker representation that minimizes the mean squared error between the target and generated speech parameters.

The STRAIGHT vocoder [30] was employed to extract 40 dimensional mel-cepstral coefficients, 10 band-aperiodicities, log F0, and U/V at 5 ms steps. The mel-cepstral coefficients were normalized to have zero-mean unit-variance during the training. In the conversion stage, the 1st-through-39th mel-cepstral coefficients and their dynamic features were converted by the VAE-based VC models. The input 0th mel-cepstral coefficients were directly used as those of target speech. The maximum likelihood parameter generation [31] was performed to generate static mel-cepstral coefficients considering their
4.2. Objective Evaluation

We calculated mel-cepstral distortions (MCDs) between target and converted mel-cepstral coefficients. The frame length of these mel-cepstral coefficients was aligned by using the DTW algorithm. We also investigated the effects produced by the number of utterances used for training one-to-one VC models or estimating speaker representations in the proposed VAE-based many-to-many VC. The four VC models used in one-to-one VC were trained by using 5, 10, 25, 50, 100, and 200 utterances. In many-to-many VC, speaker representations for the target speakers were estimated by using the same numbers of utterances as used in one-to-one VC. The MCDs were averaged over all of the possible VC settings in each of the “m2m,” “m2f,” “f2m,” and “f2f” conversion.

Figure 3 shows the evaluation results we obtained for one-to-one VC. Although the VC models were trained with completely non-parallel speech corpora, the MCDs of the proposed “VAE-SC-PPG” and “VAE-DV-PPG” were significantly improved compared with the conventional “VAE-SC” and became closer to those of “FFNN” trained with fully-parallel speech corpora. Moreover, we observed that the proposed methods even decreased the MCDs better than “FFNN” when the number of the training utterances was small, as shown in Figs. 3(b) and 3(d). One possible reason is misalignment by the DTW-based pre-processing, i.e., “FFNN” requires the DTW to align features, but our method does not. In addition, the MCDs of “VAE-DV-PPG” were slightly lower than those of “VAE-SC-PPG,” which suggested that the continuous speaker representations worked better in the VAE-based non-parallel VC.

Figure 4 shows the evaluation results obtained for the proposed VAE-based non-parallel and many-to-many VC. In all of the VC settings, we found that the MCDs of “VAE-DV-PPG” were always lower than those of “VAE-SC-PPG,” regardless of the number of utterances used for the speaker adaptation. These results indicate that using d-vectors is more effective than adapting speaker codes in the VAE-based non-parallel and many-to-many VC.

4.3. Subjective Evaluation

We conducted subjective evaluations on the naturalness and speaker similarity of the converted speech. We simultaneously compared six VC models: the two conventional models (“FFNN” and “VAE-SC”) and the two proposed models (“VAE-SC-PPG” and “VAE-DV-PPG”) for both of the one-to-one and many-to-many VC settings. The fully-parallel 400 utterances of the source and target speakers were only used in “FFNN.” Their non-parallel 200 utterances were used in the three VAE-based VC.
models for one-to-one VC. 100 target speaker utterances were used for estimating speaker representations of the target speakers in the two proposed models for many-to-many VC. Five-point scale mean opinion scores (MOS) tests were conducted for evaluating naturalness. Speech samples generated by each model were presented to listeners in random order. Similarly, five-point scale differential MOS (DMOS) tests were conducted for evaluating speaker similarity. The reference samples were re-synthesized speech presented with corresponding converted speech in random order. Fifty listeners participated in each of the evaluations for “m2m,” “m2f,” “f2m,” and “f2f” VC by using our crowd-sourced evaluation systems. Each listener evaluated the naturalness or speaker similarity of 30 converted speech samples randomly selected from all possible combinations of the test utterances and target speakers, and the total number of listeners was 400. Similar to the objective evaluation in Sect. 4.2, the results of the MOS or DMOS tests were averaged over all of the possible VC settings.

Table 1 shows the results obtained. In this evaluation, only “FFNN” was trained with fully-parallel utterances and the scores were referred to the ideal baseline of the VC models. Focusing on the one-to-one VC scores, we found the two proposed VC models achieved significantly higher scores than those of the conventional “VAE-SC” for both naturalness and speaker similarity. This demonstrated that the PPGs predicted by the pretrained ASR models successfully improved the converted speech quality in the VAE-based non-parallel VC. On the other hand, focusing on the many-to-many VC scores, we observed that the MOS and DMOS of the two proposed models were lower than those in one-to-one VC setting, although they were still superior to the conventional “VAE-SC.” This is reasonable because the speakers are unseen in the training of the VC models for many-to-many settings but are seen in that for one-to-one settings. These results suggest that the conventional VAE-based non-parallel VC can be extended to many-to-many VC. We also found that the speaker code adaptation was effective in improving the naturalness, while the use of d-vectors was effective in improving the speaker similarity of converted speech in many-to-many VC.

4.4. Effects of Training Data and Dimensionality of d-vectors in Proposed VAE-based VC

We also investigated the effects of hyperparameters in the proposed VAE-based non-parallel and many-to-many VC. We changed the number of speakers used for training the ASR, ASV and VAE-based VC models with three settings: 25 males and 25 females (“50spk”), 65 males and 65 females (“130spk”), and 130 males and 130 females (“260spk”). We also changed the dimensionality of d-vectors with six settings: 1 (“1d”), 2 (“2d”), 4 (“4d”), 8 (“8d”), 16 (“16d”), and 32 (“32d”). In total, there were 18 settings for building our VC models.
4.4.1. Evaluation of d-vector dimensionality

We investigated the effects of d-vector dimensionality in the converted speech quality of the proposed VC. Firstly, we fixed the number of speakers at 260, and compared the quality of speech samples converted by the proposed VC models with the six different settings for the d-vector dimensionality (i.e., “1d,” “2d,” “4d,” “8d,” “16d,” and “32d,”). Similar to Sect. 4.3, we conducted a series of MOS and DMOS tests for the evaluation of naturalness and speaker similarity of converted speech, respectively. Fifty listeners participated in each of the evaluation by using our crowd-sourced evaluation systems, and the total number of listeners was 400.

Table 2 shows the results obtained. From Table 2(b), we observed that the speaker similarity was significantly deteriorated when we used the d-vector dimensionality lower than eight. This result is consistent with the ASV performance described in Appendix of this paper. These results suggest that the use of inaccurate d-vectors for distinguishing speaker individuality degrades the speaker similarity of converted speech. Henceforth, the settings of lower d-vector dimensionality, i.e., “1d,” “2d,” and “4d,” were omitted from the following pairwise comparisons.

Secondly, to find optimal settings of the d-vector dimensionality, we fixed the number of speakers at 50, 130, or 260, and compared a pair of converted speech samples produced from the proposed VAE-based VC models with the three different d-vector dimensionality settings (i.e., “8d,” “16d,” and “32d,”). We conducted preference AB tests for naturalness of converted speech, and preference XAB tests for speaker similarity of converted speech. Twenty five listeners participated in each of the evaluations using our crowd-sourced evaluation systems, and the total number of listeners was 1,800. Each listener evaluated 10 samples.

Tables 3 and 4 show respectively subjective evaluation results obtained for converted speech naturalness and speaker similarity. From the results, we found two noteworthy points. When the number of speaker was small, the use of the higher d-vector dimensionality (i.e., “16d” and “32d”) significantly improved both the naturalness and speaker similarity in most cases. In contrast, these tendencies were not observed when we used the largest

| Table 2 | Subjective evaluation results for (a) naturalness and (b) speaker similarity of speech converted by proposed VAE-based VC models with different d-vector dimensionality and their 95% confidence intervals. The bold and underlined scores mean the highest and lowest ones among the six settings, respectively. |
|---------|-------------------|-------------------|-------------------|
| (a) Naturalness | m2m | m2f | f2m | f2f |
| 1d | 2.60 ± 0.12 | 2.63 ± 0.13 | 2.67 ± 0.14 | 2.80 ± 0.14 |
| 2d | 2.71 ± 0.12 | 2.64 ± 0.13 | 2.67 ± 0.14 | 2.80 ± 0.14 |
| 4d | 2.80 ± 0.13 | 2.58 ± 0.13 | 2.84 ± 0.14 | 2.75 ± 0.14 |
| 8d | 2.79 ± 0.13 | 2.60 ± 0.13 | 2.75 ± 0.12 | 2.83 ± 0.14 |
| 16d | 2.85 ± 0.13 | 2.61 ± 0.14 | 2.89 ± 0.13 | 2.86 ± 0.14 |
| 32d | 2.94 ± 0.14 | 2.64 ± 0.14 | 2.83 ± 0.13 | 2.84 ± 0.14 |
| (b) Speaker similarity | m2m | m2f | f2m | f2f |
| 1d | 2.14 ± 0.13 | 2.32 ± 0.14 | 2.01 ± 0.13 | 2.11 ± 0.14 |
| 2d | 2.18 ± 0.14 | 2.26 ± 0.13 | 2.01 ± 0.13 | 2.12 ± 0.13 |
| 4d | 2.39 ± 0.15 | 2.19 ± 0.13 | 2.40 ± 0.14 | 2.07 ± 0.13 |
| 8d | 2.44 ± 0.14 | 2.42 ± 0.14 | 2.45 ± 0.13 | 2.18 ± 0.13 |
| 16d | 2.41 ± 0.14 | 2.39 ± 0.13 | 2.36 ± 0.14 | 2.25 ± 0.14 |
| 32d | 2.49 ± 0.14 | 2.47 ± 0.14 | 2.43 ± 0.15 | 2.22 ± 0.13 |

| Table 3 | Subjective evaluation results for converted speech naturalness with fixed numbers of speakers and varied d-vector dimensionality. Bold indicates the more preferred method with p-value < 0.05. |
|---------|-------------------|-------------------|
| spk | 8d vs. 16d | 16d vs. 32d | 8d vs. 32d |
| 50 | m2m | 0.315–0.685 | 0.577–0.423 | 0.396–0.604 |
| m2f | 0.442–0.558 | 0.444–0.556 | 0.337–0.663 |
| f2m | 0.337–0.663 | 0.524–0.476 | 0.384–0.616 |
| f2f | 0.368–0.632 | 0.464–0.536 | 0.304–0.696 |
| 130 | m2m | 0.468–0.532 | 0.600–0.400 | 0.428–0.572 |
| m2f | 0.492–0.508 | 0.664–0.336 | 0.538–0.462 |
| f2m | 0.396–0.604 | 0.664–0.336 | 0.415–0.585 |
| f2f | 0.512–0.488 | 0.692–0.308 | 0.568–0.432 |
| 260 | m2m | 0.512–0.488 | 0.468–0.532 | 0.532–0.468 |
| m2f | 0.500–0.500 | 0.516–0.484 | 0.532–0.468 |
| f2m | 0.472–0.528 | 0.492–0.502 | 0.552–0.448 |
| f2f | 0.524–0.476 | 0.556–0.444 | 0.504–0.496 |

| Table 4 | Subjective evaluation results for speaker similarity of converted speech with fixed numbers of speakers and varied d-vector dimensionality. Bold indicates the more preferred method with p-value < 0.05. |
|---------|-------------------|-------------------|
| spk | 8d vs. 16d | 16d vs. 32d | 8d vs. 32d |
| 50 | m2m | 0.450–0.550 | 0.485–0.515 | 0.528–0.472 |
| m2f | 0.472–0.528 | 0.472–0.528 | 0.476–0.524 |
| f2m | 0.420–0.580 | 0.468–0.532 | 0.412–0.588 |
| f2f | 0.462–0.538 | 0.496–0.504 | 0.438–0.562 |
| 130 | m2m | 0.428–0.572 | 0.508–0.492 | 0.454–0.546 |
| m2f | 0.548–0.452 | 0.480–0.520 | 0.528–0.472 |
| f2m | 0.484–0.516 | 0.504–0.496 | 0.500–0.500 |
| f2f | 0.588–0.412 | 0.476–0.524 | 0.560–0.440 |
| 260 | m2m | 0.452–0.548 | 0.527–0.473 | 0.516–0.484 |
| m2f | 0.488–0.512 | 0.536–0.464 | 0.460–0.540 |
| f2m | 0.464–0.536 | 0.540–0.460 | 0.432–0.568 |
| f2f | 0.508–0.492 | 0.464–0.536 | 0.569–0.431 |
number of speakers (i.e., “260spk”), and “8d” even achieved higher speaker similarity than “32d.” These results suggest that there is a trade-off between the $d$-vector dimensionality and number of speakers, i.e., if we want to reduce the $d$-vector dimensionality, the number of speakers should be large enough to deal with unseen speakers in the VC process.

4.4.2. Evaluation of number of speakers

Thirdly, to investigate the effects produced by the number of speakers, we fixed the dimensionality of $d$-vectors whose preference scores shown in Tables 3 and 4 were the best among the different settings and compared converted speech quality. Tables 5 and 6 show respectively the subjective evaluation results obtained for naturalness and speaker similarity. We observed that an increased number of speakers tended to improve both the naturalness and speaker similarity of the converted speech. To investigate the reason, we calculated root mean square errors (RMSEs) between two PPGs predicted from source and target speakers. The DTW algorithm was performed to align frame lengths of PPGs. Table 7 shows the results, and we found that the RMSEs decreased in proportion to the number of speakers, which may account for the improvement in converted speech quality.

4.4.3. Visualization of latent variables

Finally, to explore how informative the $d$-vectors and latent variables are, we made scatter plots of them using principal component analysis. In the following visualization, the $d$-vectors or latent variables are averaged over all evaluation data. Bold indicates the lowest RMSE in each row.

### Table 5
Subjective evaluation results for naturalness of converted speech with varied numbers of speakers and the best $d$-vector dimensionality for each number of speakers. Here, we used respectively “32d” for “50spk,” “16d” for “130spk,” and “8d” for “260spk.” Bold indicates the more preferred method with $p$-value < 0.05.

|       | 50spk vs. 130spk | 130spk vs. 260spk | 50spk vs. 260spk |
|-------|------------------|-------------------|------------------|
| m2m   | 0.316–0.684      | 0.460–0.540       | 0.240–0.760      |
| m2f   | 0.476–0.524      | 0.408–0.592       | 0.440–0.560      |
| f2m   | 0.364–0.636      | 0.438–0.562       | 0.324–0.676      |
| f2f   | 0.496–0.504      | 0.492–0.508       | 0.473–0.527      |

### Table 6
Subjective evaluation results for speaker similarity of converted speech with varied numbers of speakers and the best $d$-vector dimensionality for each number of speakers. Here, we used respectively “32d” for “50spk,” “16d” for “130spk,” and “8d” for “260spk.” Bold indicates the more preferred method with $p$-value < 0.05.

|       | 50spk vs. 130spk | 130spk vs. 260spk | 50spk vs. 260spk |
|-------|------------------|-------------------|------------------|
| m2m   | 0.436–0.564      | 0.476–0.524       | 0.285–0.715      |
| m2f   | 0.496–0.504      | 0.548–0.460       | 0.416–0.584      |
| f2m   | 0.408–0.592      | 0.427–0.573       | 0.360–0.640      |
| f2f   | 0.462–0.538      | 0.532–0.468       | 0.448–0.552      |

### Table 7
RMSEs of PPGs predicted from source and target speakers averaged over all evaluation data. Bold indicates the lowest RMSE in each row.

|       | 50spk | 130spk | 260spk |
|-------|-------|--------|--------|
| m2m   | 0.0564| 0.0534 | **0.0539** |
| m2f   | 0.0566| 0.0542 | **0.0541** |
| f2m   | 0.0566| 0.0542 | **0.0541** |
| f2f   | 0.0563| 0.0541 | **0.0539** |

---

**Fig. 5** Scatter plots of $d$-vectors compressed with principal component analysis. Each point denotes representation of one speaker, and the three male and three female speakers used in experimental evaluations were marked with texts “M” and “F” in this plot, respectively. To place males’ cluster in the left side and females’ cluster in the right side, some figures were rotated 180 degrees.
proposed VAE-based VC, rather that the space constructed for the better ASV performance.

Figure 6 shows the visualization of latent variables of VAEs. The “CSM” in this plot means the class separability measure calculated as the trace of the between-class scatter matrix divided by that of the within-class scatter matrix [34]. Since we calculated the CSM values in the latent variables space regarding speakers as classes, the numerator and denominator of the CSM definition should represent inter-speaker variation and inter-phonetics variation, respectively. We found that the resulting SCM values were always greater than 1.0 as shown in Fig. 6, which indicates that the inter-speaker variation is dominant in the latent variables. In fact, some of the six test speakers tend to position out of the distribution of the latent variables when we used the lower number of speakers for the training (e.g., “M1” and “M3” in Fig. 6(c)(1)). These results suggest that the latent variables of the proposed VAE-based VC models quantify how well the VC models fit the speaker individuality, rather than phonetic contents [11].

5. CONCLUSION

We proposed novel non-parallel and many-to-many voice conversion (VC) using variational autoencoders (VAEs). In the proposed VC, pretrained automatic speech recognition (ASR) models are introduced into the VAE-based VC for improving converted speech quality by using phonetic posteriorgrams (PPGs) predicted by the ASR models. To estimate speaker representations for unseen speakers, pretrained automatic speaker verification (ASV) models are employed to extract d-vectors from a bottleneck layer of the ASV models. Experimental results demonstrated that PPGs significantly improved both naturalness and speaker similarity of the converted speech compared to conventional VAE-based VC and that d-vectors were effective speaker representations for the VAE-based non-parallel and many-to-many VC. We also investigated the effects of hyperparameters of the proposed VC, and found that 1) a large d-vector dimensionality that gives the better ASV performance does not necessarily improves converted speech quality, and 2) a large number of pre-stored speakers tends to improve converted speech quality. In the future, we will further investigate the effects of the difference between the source and target speakers (e.g., the MCD or the distance in the d-vector space), since it should affect the conversion accuracy in many-to-many VC. In addition, we will devise an effective way for building our VC models with less amount of the training data (e.g., semi-supervised learning or active learning).

ACKNOWLEDGEMENTS

Part of this work was supported by JSPS KAKENHI Grant Number 18J22090.

REFERENCES

[1] S. Aryal and R. G. Osuna, “Can voice conversion be used to reduce non-native accents?” Proc. ICASSP 2014, pp. 7929–7933 (2014).
[2] H. Doi, T. Toda, T. Nakano, M. Goto and S. Nakamura, “Singing voice conversion method based on many-to-many Eigenvoice conversion and training data generation using a singing-to-singing synthesis system,” Proc. APSIPA ASC, pp. 1–6 (2012).
[3] A. B. Kain, J.-P. Hosom, X. Niu, J. P. H. van Santen, M. Fried- Oken and J. Staehely, “Improving the intelligibility of dysarthric speech,” Speech Commun., 49, 743–756 (2007).
[4] Y. Stylianou, O. Cappé and E. Moulines, “Continuous probabilistic transform for voice conversion,” IEEE Trans. Speech Audio Process., 6, 131–142 (1998).
[5] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black and K. Prahallad, “Voice conversion using artificial neural networks,” Proc. ICASSP 2009, pp. 3893–3896 (2009).
[6] T. Toda, A. W. Black and K. Tokuda, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” IEEE Trans. Audio Speech Lang. Process., 15, 2222–2235 (2007).
[7] G. Hinton, L. Deng, D. Yu, G. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” IEEE Signal Process. Mag., 29, 82–97 (2012).
[8] E. Variani, X. Lei, E. McDermott, I. L. Moreno and J.
Gonzalez-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” Proc. ICASSP 2014, pp. 4080–4084 (2014).

[9] Y. Saito, S. Takamichi and H. Saruwatari, “Statistical parametric speech synthesis incorporating generative adversarial networks,” IEEE/ACM Trans. Audio Speech Lang. Process., 26, 84–96 (2018).

[10] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” arXiv, vol. abs/1312.6114 (2013).

[11] 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,” Proc. APSIPA ASC, pp. 1–6 (2016).

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, “Generative adversarial nets,” Proc. NIPS 2014, pp. 2672–2680 (2014).

[13] J.-Y. Zhu, T. Park, P. Isola and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” Proc. ICCV 2017, pp. 2223–2232 (2017).

[14] T. Kaneko and H. Kameoka, “CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks,” Proc. EUSIPCO 2018, pp. 2114–2118 (2018).

[15] F. Fang, J. Yamagishi, I. Echizen and J. Lorenzo-Trueba, “High-quality non-parallel voice conversion based on cycle-consistent adversarial network,” Proc. ICASSP 2018, pp. 5279–5283 (2018).

[16] N. Hojo, Y. Iijima and H. Mizuno, “DNN-based speech synthesis using speaker codes,” IEICE Trans. Inf. Syst., E101-D, 462–472 (2018).

[17] 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).

[18] 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).

[19] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” Proc. ICASSP 2000, pp. 1315–1318 (2000).

[20] J. Duchi, E. Hazan and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” J. Mach. Learn. Res., 12, 2121–2159 (2011).

[21] X. Glorot, A. Bordes and Y. Bengio, “Deep sparse rectifier neural networks,” Proc. AISTATS 2011, pp. 315–323 (2011).

[22] M. W. Wolfel, “Channel selection by class separability measures for automatic transcriptions on distant microphones,” Proc. Interspeech 2007, pp. 582–585 (2007).

APPENDIX: PERFORMANCE OF ASR AND ASV MODELS

We investigated the ASR and ASV performances under several settings of two hyperparameters: the number of speakers and d-vector dimensionality. Using the utterances of six speakers (three males and three females for the VC evaluation), we measured two performances: the frame-wise phoneme error rate (PER) of ASR and equal error rate (EER) of ASV. For measuring the EER, we used 50 utterances of each speaker for enrollment (i.e., extracting d-vectors) and other 25 utterances for evaluation (i.e., computing the cosine distance between input and claimed speaker’s d-vectors). We applied the L2 normalization to the d-vectors [8].

Table A-1 shows the results obtained. We observed that the PER significantly increased under the setting of the smallest number of speakers (i.e., “50spk”), which may affect the naturalness of converted speech as we conclude

| PER [%] | EER [%] | 1d | 2d | 4d | 8d | 16d | 32d |
|--------|--------|----|----|----|----|-----|-----|
| 50spk  |       | 54.2 | N/A | 29.5 | 10.4 | 6.96 | 2.97 | 2.48 |
| 130spk |       | 48.6 | N/A | 29.1 | 13.1 | 7.81 | 1.60 | 1.33 |
| 260spk |       | 49.2 | N/A | 27.4 | 10.3 | 6.22 | 4.07 | 0.78 |
in Sect. 4.4.2. We also revealed that the EER drastically increased under the settings of the smaller $d$-vector dimensionality (i.e., “1d,” “2d,” and “4d”), which may affect the speaker similarity of converted speech as we concluded in Sect. 4.4.1.

Yuki Saito received the M.S. degree from the Graduate School of Information Science and Technology, The University of Tokyo, Japan, in 2018. He is currently a Ph.D. student at The University of Tokyo. His research interests include speech synthesis, voice conversion, and machine learning. He has received eight paper awards including The 34th TELECOM System Technology Award for Students from TAF. He is a Student Member the Acoustical Society of Japan (ASJ) and a Student Member of IEEE Signal Processing Society (SPS).

Taiki Nakamura is presently an undergraduate student in the Faculty of Engineering at the University of Tokyo, Japan. He was engaged in this research as an intern student. Since then, he has worked part-time as a speech recognition researcher for enterprises. He is a Student Member the Acoustical Society of Japan (ASJ).

Yusuke Ijima received the B.E. degree in electric and electronics engineering from National Institution for Academic Degrees and University Evaluation by graduation from Yatsushiro National College of Technology, Japan, in 2007, and the M.E. and Ph.D. degrees in information processing from Tokyo Institute of Technology, Japan, in 2009 and 2015, respectively. He joined NTT Cyber Space Laboratories (currently Media Intelligence Laboratories) in 2009, where he engaged in the research and development of speech synthesis. He is a member of the Acoustical Society of Japan (ASJ), the International Speech Communication Association (ISCA) and the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.

Kyosuke Nishida received the B.E., M.I.S., and Ph.D. in information science and technology from Hokkaido University, Japan, in 2004, 2006, and 2008. He joined NTT in 2009 and is presently a Distinguished Researcher of NTT Media Intelligence Laboratories. He received the Yamashita SIG Research Award from the Information Processing Society of Japan (IPSJ) in 2015 and the Kambayashi Young Researcher Award from the Database Society of Japan (DBSJ) in 2017. His current interests include natural language processing and artificial intelligence. He is a member of the Association for Computing Machinery (ACM), IPSJ, the Association for Natural Language Processing (NLP), the Institute of Electronics, Information and Communication Engineers (IEICE), and DBSJ.

Shinnosuke Takamichi received the Ph.D. degree from the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Japan, in 2016. He is currently an Assistant Professor at The University of Tokyo. He has received more than ten paper/achievement awards including the 3rd IEEE Signal Processing Society Japan Young Author Best Paper Award.