Vol.2019-MUS-123 No.58
Vol.2019-SLP-127 No.58
2019/6/23

Non-parallel Voice Conversion with Controllable Speaker Individuality using Variational Autoencoder

Tuan Vu Ho[1,a)] Masato Akagi[1,b)]

Abstract: This paper investigates a voice conversion (VC) system that can both perform speaker adaptation and control voice characteristics. To achieve this goal, we formulate the voice conversion task as learning the disentanglement of speaker-related information and linguistic information by a variant of Variational Autoencoder (VAE), which we shall call Cycle-consistent Variational Autoencoder (CycleVAE). Neither parallel training utterances, linguistic label nor time alignment procedure is required to train our system. After training on utterances from many speakers, our proposed VC system can adapt to arbitrary target voice using only one reference utterance from target speaker. By interpolating the discovered speaker embedding vector that represent voice characteristics, our proposed VC system can synthesize new voices without any reference target utterance , which makes it beneficial for many practical applications. The preliminary subjective evaluation of non-parallel voice conversion task shows that our proposed system obtains higher naturalness and comparable speaker similarity than the conventional VC using look-up one-hot encoded speaker vector.

Keywords: Voice conversion, voice control, variational autoencoder, non-parallel data

1. INTRODUCTION

Voice conversion (VC) is a special type of voice transformation (VT) whose aim is to manipulating speaker characteristics in the speech signal while preserving linguistic information [1]. This technique is beneficial in many practical applications such as intelligibility enhancement for speech disorder patients, or enhancing Human-Machine Interface experience. VC approach can be categorized into 2 groups: rule-based approaches and statistical approaches.

Rule-based approaches aims to modify acoustic features that correspond to the speaker individuality such as fundamental frequency (F_0) and formants by some manually derived rules. However, since different rules must be applied for different speakers, these approaches are impractical and less preferred than statistical approach.

On the other hand, statistical approaches use machine learning technique to modify the acoustic features. These approaches are more flexible to adapt to new speaker than rule-based method. The most straight-forward statistical approach for VC is to perform mapping from source acoustic features to target acoustic features. This approach requires a parallel training data, in which the source and target utterances contain identical linguistic information so that the differences in speaker voice characteristics could be learned. The conventional method for this approach is using Gaussian Mixture Model (GMM) to model the joint probability of source and target acoustic features [2]. However, synthesized speech using GMM-based method often suffered from over-smoothing degradation. Therefore, lately, Deep Neural Network (DNN) has been employed to perform the mapping task. With sufficient training data, DNN-based model outperforms GMM-based model in both speech naturalness and target-similarity.

Despite the simplicity of mapping approach, parallel training data is often expensive to obtain. Therefore, a new set of method that can perform speaker adaptation using non-parallel data is investigated. The earliest non-parallel VC method was proposed by Toda et al [2], in which a Eigen GMM-based model is used to describe speaker characteristic by a set of base speaker. However, although the speaker adaptation phase can work with non-parallel data, it requires parallel-data in the training phase. Later, various methods were proposed that can use non-parallel data in both training phase and adaptation

[1] Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1211, Japan
[a)] tuanvu.ho@jaist.ac.jp
[b)] akagi@jaist.ac.jp

phase. Some of the most popular methods are Restricted Boltzmann machine (RBM), Variational Autoencoder (VAE), and Generative Adversarial Network (GAN). All these three methods share the same principle of disentangling speaker-related information and linguistic information from speech waveform.

However, most prior non-parallel VC methods only focus on categorized speaker adaptation since a target voice is required as a reference to perform voice conversion. In other words, controllability of the degree of speaker individuality has not been much interested. These limitations restricted the usefulness of VC system in some situations, such as in a story teller system, when collecting utterances from a large number of target voices is unrealistic. In this situation, the VC system with the controllable voice characteristics is desirable as it can freely manipulate the source voice to generate any new fictitious voice without the recordings from the target speakers. Moreover, most VC model require retraining when adapting to an unseen-target speaker. The controllability can also avoid this problem as the VC model can synthesize waveform with the desired voice characteristics extracted from the reference utterance. This controllability is also beneficial in many other voice transformation fields such as emotional voice conversion, voice dubbing in movie post-production, creating new voices for text-to-speech system, speech enhancement, and voice editing software.

To achieve this goal, we propose a new VC framework based on a variant of VAE, which we call CycleVAE, that can simultaneously disentangle speaker-related information with linguistic information and discover the latent structure of speaker characteristic. After training on a multi-speaker dataset, a speaker embedding vector that represents voice characteristics is obtained. By manipulating the speaker embedding vector, we can obtained the synthesized waveform with desired voice characteristics.

The significant of our proposed VC system are:

- Control the characteristics of synthesized voice using non-parallel training data.
- Can perform speaker adaptation using a minimum of one utterance from target speaker.
- Can convert waveform from both seen- and unseen-source speaker to unseen-target speaker and fictitious speaker.

2. Voice Conversion with Variational Autoencoder

Proposed by Kingma et al. and Rezende et al. [6], VAE is a powerful probabilistic model that can uncover the latent structure of the data. Without much modification, VAE can be easily apply in VC tasks [7].

Assume that the latent variable $\mathbf{Z}$ represent the linguistic information conveyed in acoustic features X follows normal distribution $\mathcal{N}(\mathbf{0},\mathbf{I})$ that independent with the speaker information. The encoder part of VAE estimates the posterior $p_\theta(Z|X) = \mathcal{N}(\mu(X),\sigma(X))$. Then the latent variable Z is sampled from the posterior as $z \sim p(Z|X)$. However, back-propagation is impossible if Z is directly sampled from the posterior $p_\theta(Z|X)$. Therefore, reparaterization trick is applied by sampling an independent variable ε from normal distribution $\mathcal{N}(\mathbf{0},\mathbf{I})$ and then performing scale and shift operation. In summary, the procedure of estimating latent variable Z is as follows:

$$\begin{aligned}\mu &= f_{enc\_\mu}(\mathbf{X})\\ \sigma &= f_{enc\_\sigma}(\mathbf{X})\\ \varepsilon &\sim \mathcal{N}(\mathbf{0},\mathbf{I})\\ \mathbf{Z} &= \mu + \sigma \circ \varepsilon\end{aligned} \tag{1}$$

To reconstruct the input acoustic feature X, beside the linguistic information in latent variable Z, additional variable y that contains speaker information is introduced. The variable y can be expressed as a one-hot encoded vector that represents speaker identity. From variable Z and y, the decoder part of the VAE then reconstruct the acoustic features X.

$$\overline{X} = f_{dec}(Z,Y) \tag{2}$$

The encoder and decoder are jointly trained by maximizing the objective function defined as:

$$\mathscr{L}_{obj} = D_{KL}(p_\theta(z|x)||p(z)) + \mathbb{E}_{z\sim p_\theta(z|x)}(p(x|z)), \tag{3}$$

where D_{KL} is the Kullback-Leibler divergence between the estimated posterior $p_\theta(z|x,y)$ and the true prior distribution $p(z)$. Since $p(z)$ is assumed to follow normal distribution, the D_{KL} can be expressed in closed form as:

$$D_{KL}(p_\theta(z|x)||p(z)) = -\frac{1}{2}\sum(1+\log\sigma^2-\mu^2+\sigma^2) \tag{4}$$

The second term in the RHS of Eq. 3 is the reconstruction loss. Assuming that the acoustic feature X also follows Gaussian distribution, the term $\mathbb{E}_{z\sim p_\theta(z|x)}(p(x|z,y))$ can be described by a simple mean-square difference between reconstructed acoustic feature and original acoustic feature.

$$\mathbb{E}_{z\sim p_\theta(z|x)}(p(x|z)) = -\frac{1}{2}\sum(\overline{X}-X)^2 \tag{5}$$

3. Proposed Method

3.1 Infer speaker embedding using back-propagation

In conventional VAE-based VC, speaker identity is represented as a one-hot vector. However, this type of encoding does

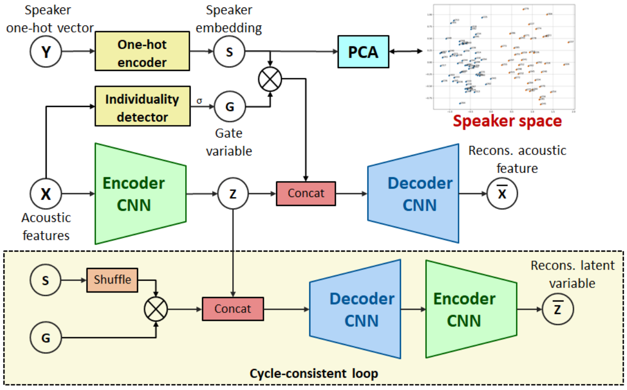

Fig. 1 Overview of proposed VC system

not include any other information of the speaker's voice characteristics such as gender or age. To overcome this problem, we use different interpretation of speaker identity by letting the model self-derived the most suitable speaker embedding during training process. Let y is the one-hot vector represent speaker identity, the speaker embedding vector s is:

$$s = \mathbf{W}\cdot y^{\mathsf{T}} + \mathbf{B}, \tag{6}$$

where $\mathbf{W}$ and $\mathbf{B}$ is a learnable kernel and bias in a fully-connected NN layer. In this interpretation, the one-hot encoded vector y acts as a switch to select correspond row vector in matrix $\mathbf{W}$. With this interpretation, 2 speakers with alike voice characteristics may have very similar speaker embedding.

This interpretation can be expanded into by adding more layer and applying non-linear activation such as tanh or sigmoid. In this case, the speaker embedding s is

$$s = \mathbf{W_n}\cdot \ldots f(\mathbf{W_1}\cdot f(\mathbf{W_0}\cdot y^{\mathsf{T}} + \mathbf{B_0}) + \mathbf{B_1})\ldots + \mathbf{B_n}, \tag{7}$$

where f is a non-linear function. Although this interpretation is convenient to explain voice characteristics, however, the speaker embedding is only available for speakers in the training set. Therefore, to perform voice conversion on new speaker that non in the training set, an additional classifier is used to map from acoustic features to speaker embedding vector. After the classifier is trained, a speaker embedding vector from new speaker can be estimated using only a few seconds of recording.

3.2 Cycle-consistent loss

In our proposed system, we also introduced some additional constrained to improve the naturalness of the synthesized speech and enhance the disentanglement of the latent variable z and speaker information s.

As the latent variable z is assumed to be independent with among speaker information, we propose to use the cycle-consistent loss, which aims to ensure the invariant of z when changing the speaker information. To achieve this goal, the speaker identity vector is shuffled before inputting to the decoder. Then the estimated mean of latent variable $\overline{\mu}$ of the synthesized acoustic feature is calculated. To enhance the invariant of latent variable, additional penalty on the difference between $\overline{\mu}$ and μ is introduced, which we call cycle-consistent loss.

3.3 Modulation loss

To improve the naturalness of the synthesized speech, we also incorporate the Modulation Spectrum (MS) loss in the proposed model because of its beneficial effect on speech naturalness. Similar to [8], the MS of parameter sequence x is defined as follows:

$$\begin{aligned} \mathbf{s}(\mathbf{X}) &= \left[\mathbf{s}(1)^{\top},\cdots,\mathbf{s}(d)^{\top},\cdots,\mathbf{s}(D)^{\top}\right] \\ s(d) &= [s_d(0),\cdots,s_d(f),\cdots,s(D_s)] \\ s_d(f) &= abs(FFT(\mathbf{x}(d))) \end{aligned} \tag{8}$$

The modified log-likelihood function for the VAE model considering the modulation spectrum is defined as follow:

$$\begin{aligned} \overline{L}_{ms}(\theta,\phi;\mathbf{x}_n) &= -D_{KL}(q_\phi(\overline{\mathbf{z}}_n|\mathbf{x}_n)||p(\mathbf{z}_n)) \\ +log\ p_\theta(\mathbf{x}_n|\overline{\mathbf{z}}_n,\mathbf{y}_n) &+ w.log\ p(s(\mathbf{x})|\overline{\mathbf{z}}_n,\mathbf{A}^{(X)}) \end{aligned} \tag{9}$$

The final term in Eq. 9 explicitly constrains the model to increase the log-likelihood of the modulation spectrum conditioned on the given latent variable $\overline{\mathbf{z}}_n$ and speaker identity y_n. Furthermore, we also assume that the modulation spectrum has a Gaussian distribution with a diagonal covariance matrix: $s(x) \sim N(s(x)|s(\overline{x}), diag(\sigma_s))$. Therefore, the final log-probability term in Eq. 9 can be expressed in the following closed form:

$$\begin{aligned} &log\ p(s(\mathbf{x})|\overline{\mathbf{z}}_n,\mathbf{A}^{(X)}) = \\ &-\frac{1}{2}\sum\left(log(2\pi\sigma_s^2) + \frac{(s(\mathbf{x}) - s(\overline{\mathbf{x}}))^2}{\sigma_s^2}\right) \end{aligned} \tag{10}$$

3.4 Network architecture

Figure 1 illustrates an overview of our proposed VC model. The encoder and decoder network utilize the multi-scale CNN architecture[9] as shown in Fig. 2.

In addition to the basic VAE framework, the auxiliary gate variable g is introduced to control amount of the speaker individuality in the output features. The reason for this controlling is that some speech segments, such as silence, may not contains any speaker individuality. By introducing the gate variable, the model can ignore these segments by outputting the gate variable $g = 0$. The gate variable is inferred directly on the input features by a separate network.

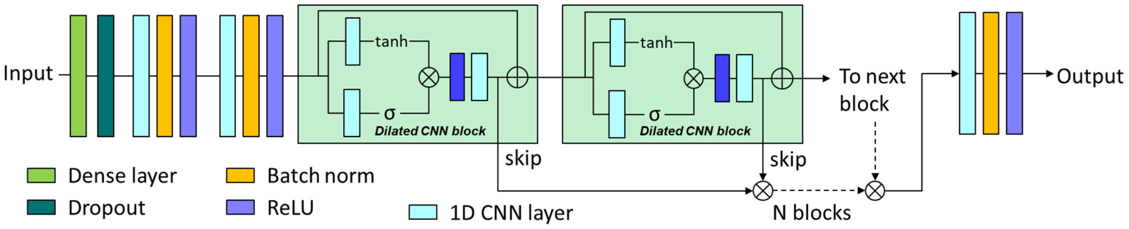

Fig. 2 Multi-scale architecture with dilated residual CNN block

4. EXPERIMENTS

4.1 Dataset

We used the VCTK corpus [10], which contains 44 hours of recordings from 109 English speakers. We divided the data into 2 subsets: training set (containing 100 speakers) and testing set. The testing set is made of 2 groups of utterances. One group contains utterances from 9 held out speakers from the training set (unseen speakers). The second group contains 2 held out utterances of each speakers from the training set (seen speakers).

As speech features, we used WORLD vocoder to extract F_0, spectral sequence, and aperiodicity from speech waveform. Then the spectral sequence is transformed to 60-order mel-cepstral coefficients (mcc). We used the 2^{nd} to 31^{th} mcc coefficients along with interpolated F_0 and voice/unvoice flag as the input features. For the rest of mcc-coefficients and aperiodicity, we keep unchanged during conversion process.

The VC model and speaker embedding model are trained separately. We first train the VC model to obtained the speaker embedding table. Then we trained the speaker embedding model to map from speech features to embedding vector. Both VC model and speaker embedding model are trained on the same training set.

We compare the proposed model to the baseline multi-speaker VAE-based VC model that use the fixed one-hot encoded speaker vector similar to [7]. However, we keep most of the model architecture identical to the proposed model. Since the baseline model cannot work with unseen target speaker, we only evaluate the baseline model in seen source to seen target and unseen source to seen target conversion scenarios. To perform voice conversion, the baseline model uses the one-hot encoded vector, while the proposed model uses the speaker embedding extracted from an 10-second utterance of the target speaker.

4.2 Speaker embedding space

After the VC model is trained, we visualize the speaker embedding space by analyzing the speaker embedding using PCA.

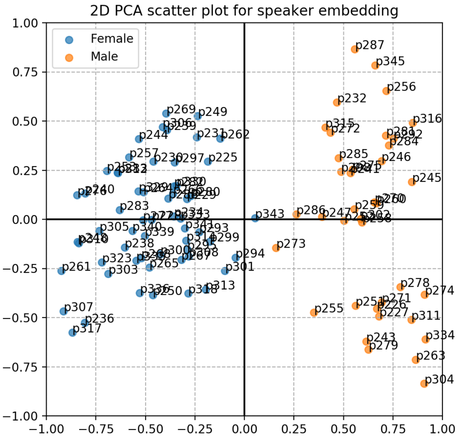

Fig. 3 Learned speaker embedding map of VCTK dataset

As shown in Fig. 3, the speakers are well separated by genders, with all female speakers lie on the left and male speakers lie on the right. This indicates that the model can learn meaningful voice characteristics of the speakers.

4.3 Fictitious speaker

We input the speaker embedding vector that is sampled from the speaker embedding space to obtain the fictitious voices that are not present in the training data. To evaluate the naturalness of the fictitious voices, we synthesized 9 utterances from a female speaker in VCTK dataset (seen speaker - p225) with the position on speaker embedding space shown in Fig. 4

4.4 Speech Naturalness

We measure the naturalness of converted speech from the baseline and proposed model using Mean-Opinion Score evaluation. Two participants (1 male, 1 female) enrolled in this preliminary test. The listeners are instructed to concentrate on the quality of the speech and rate the sample using 5 point-scale score (1: bad; 2: poor; 3: fair; 4: good; 5: excellent). The result shown in Fig. 5 indicates that the speech waveform generated from the proposed model have higher naturalness than those generated from the baseline model in all conversion scenarios (seen-source to seen-target and unseen-source to seen-target).

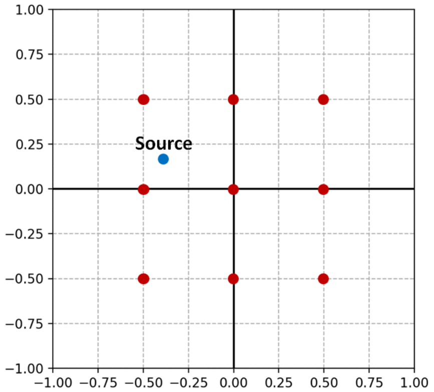

Fig. 4 **Blue**: Position of source speaker embedding vector, **Red**: Position of selected target speaker embedding vector for synthesizing fictitious voices

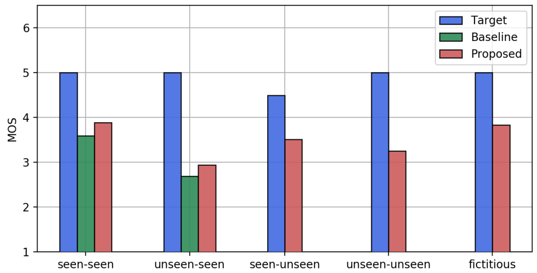

Fig. 5 Speech Naturalness average score. Higher is better.

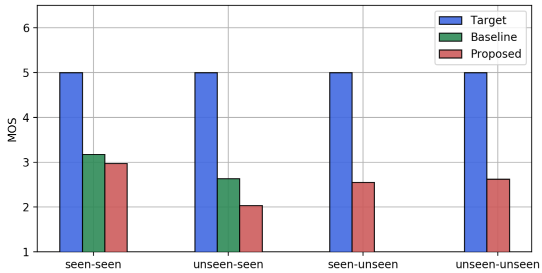

Fig. 6 Speech Similarity average score. Higher is better.

The highest MOS of the proposed model is approximate 4.0 in seen-source to seen-target conversion. Moreover, the generated speech of fictitious speakers also have very good naturalness with 3.8 MOS. For unseen target and fictitious target speaker, the naturalness of synthesized waveform is quite good with all MOS higher than 3.0.

4.5 Speech Similarity

In this experiment, the speaker similarity between the converted waveform and the target waveform is evaluated. The listeners are given a reference utterance from target speaker and several converted utterances from different source speakers. The listeners were instructed to concentrate on the voice characteristics and ignore any distortion in the stimuli. Then the listener rates the similarity between the converted utterances with the reference utterance using 5-point scale score (1: not at all similar; 2: slightly similar; 3: moderately similar; 4: very similar; 5: extremely similar). Results are shown in Fig. 6. On both seen- and unseen-source to seen-target speaker, the scores for the proposed model are lower but still comparable to the baseline model. For seen- and unseen-source to unseen-target speaker, the propose model achieves better score with around 2.7 MOS.

5. CONCLUSIONS

We have proposed a flexible VC model to deal with the challenging task of non-parallel voice conversion with controllable speaker individuality. The preliminary results show that the proposed model can synthesized speech with higher naturalness than the baseline model. Although the speaker similarity score of the proposed model is comparable to the conventional VC, there is still more room for further improvement in the future.

参考文献

[1] S. H. Mohammadi, A. Kain, “ An overview of voice conversion systems, ” Journal of Speech Communication, vol. 88, pp. 65-82.

[2] T. Toda, A. W. Black, 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, November 2007.

[3] T. Toda, Y. Ohtani, K. Shikano, “ One-to-Many and Many-to-One Voice Conversion Based on Eigenvoices, ” ICASSP, 2007.

[4] H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo,“ StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks, ” arXiv:1806.02169 [cs.SD], Jun. 2018.

[5] M. Akagi, X. Han, R. Elbarougy, Y. Hamada and J. Li, “ Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages, ” Proc. APSIPA, 2014.

[6] D. P. Kingma, M. Welling, ”Auto-Encoding Variational Bayes,” ICLR, 2014.

[7] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, Hsin-Min Wang, ”Voice Conversion from Non-parallel Corpora Using Variational Auto-encoder,”

[8] S. Takamichi, T. Toda, A. W. Black, S. Nakamura, ”Modified Post-filter to Recover Modulation Spectrum for HMM-based Speech Synthesis,” GlobalSIP, 2014.

[9] T. V. Ho, M. Akagi, ”Speech Accent and Gender Recognition using Dilated Convolution Neural Network with Skip and Residual Connection,” Acoustic Society Japan Spring Meeting, 2019.

[10] C. Veaux, J. Yamagishi, K. MacDonald, ”CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit”.