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

Tuan Vu Ho* and Masato Akagi*
* Japan Advanced Institute of Science and Technology, Japan
tuanvu.ho@jaist.ac.jp, akagi@jaist.ac.jp

***Abstract*—We propose a flexible non-parallel voice conversion (VC) system that is capable of both performing speaker adaptation and controlling speaker individuality. The proposed VC framework aims to tackle the inability to arbitrarily modify voice characteristics in the converted waveform of conventional VC model. To achieve this goal, we use the speaker embedding realized by a Variational Autoencoder (VAE) instead of one-hot encoded vectors to represent and modify the target voice's characteristics. Neither parallel training data, linguistic label nor time alignment procedure is required to train our system. After training on a multi-speaker speech database, the proposed VC system can adapt an arbitrary source speaker to any target speaker using only one sample from a target speaker. The speaker individuality of converted speech can be controlled by modifying the speaker embedding vectors; resulting in a fictitious speaker individuality. The experimental results showed that our proposed system is similar to conventional non-parallel VAE-based VC and better than the parallel Gaussian Mixture Model (GMM) in both perceived speech naturalness and speaker similarity; even when our system only uses one sample from target speaker. Moreover, our proposed system can convert a source voice to a fictitious target voice with well perceived speech naturalness of 3.1 MOS.**

***Index Terms*—Voice conversion, speaker embedding, voice characteristics control, variational autoencoder, non-parallel data**

I. 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 approaches can be categorized into 2 groups: rule-based approaches and statistical approaches.

Rule-based approaches [2]–[4] aim 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 [5]. 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 has been investigated. The first non-parallel VC method utilize an Eigen GMM-based model to describe speaker characteristic as combinations of base speakers [5]. 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 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 use of VC system in some situations, such as in a storyteller 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 requires 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 conver-

sion, 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 VAE 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 (SE) that represents voice characteristics is obtained. By manipulating the speaker embedding vector, we can obtain the synthesized waveform with desired voice characteristics. In this paper, we show that VC system using VAE with SE input (SE-VAE) has comparable performance as using VAE with one-hot encoded input (OH-VAE).

The significant of our proposed VC system are:

- Controlling the characteristics of converted voice using non-parallel training data.
- Performing speaker adaptation using a minimum of one utterance from target speaker.
- Converting waveform from both seen- and unseen-source speaker to unseen-target speaker and fictitious speaker.

II. Voice Conversion with Variational Autoencoder

Proposed by Kingma et al. and Rezende et al. [9], VAE is a powerful probabilistic model that can uncover the latent structure of the data. Previous research showed the interpolability of VAE-based latent representation [10], [11]

Assume that the latent variable $\mathbf{Z}$ represent the linguistic information conveyed in acoustic features $\mathbf{X}$ follows normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ that independent with the speaker information. The encoder part of VAE estimates the posterior $p_\theta(\mathbf{Z}|\mathbf{X}) = \mathcal{N}(\mu(\mathbf{X}), \sigma(\mathbf{X}))$. Then the latent variable $\mathbf{Z}$ is sampled from the posterior as $z \sim p(\mathbf{Z}|\mathbf{X})$. However, back-propagation is impossible if Z is directly sampled from the posterior $p_\theta(\mathbf{Z}|\mathbf{X})$. Therefore, reparameterization 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 $\mathbf{Z}$ is as follows:

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

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

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

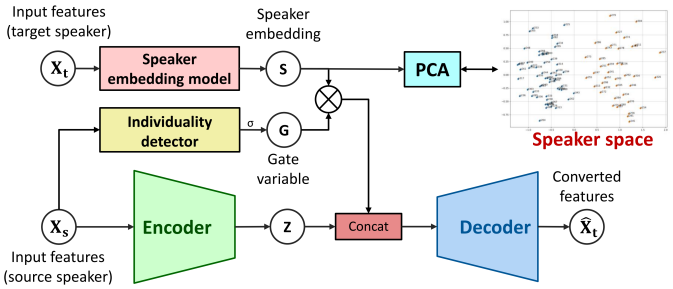

Fig. 1. Overview of proposed VC system

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

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

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 righ-hand side of Eq. 3 is the reconstruction loss. Assuming that the acoustic feature $\mathbf{X}$ 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.

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

III. Proposed Method

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

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

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 similar voice characteristics may have almost identical 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

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

where f is a non-linear function. Although this interpretation is convenient to explain voice characteristics, however, the

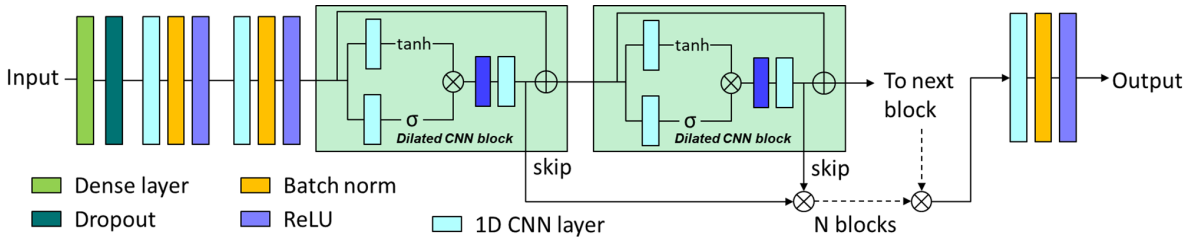

Fig. 2. Multi-scale architecture with dilated residual CNN block

speaker embedding is only available for speakers in the training set. Therefore, to perform voice conversion on a new target speaker that is not in the training set, a speaker embedding model is trained to predict the corresponding speaker embedding from acoustic features. The learned speaker embeddings obtained after training VAE model are used as the ground truth. After the speaker embedding model is trained, a speaker embedding vector from new target speaker can be estimated using only a few seconds of their recording (10 seconds in our experiments).

B. 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 [12], the MS of parameter sequence x is defined as follows:

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

where D is the number of channel of $\mathbf{x}$, D_s is the number of frame of $\mathbf{X}$, $s_d(f)$ is the FFT of channel d at frequency bin f.

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:

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

C. Network Architecture

Fig. 1 illustrates an overview of our proposed SE-VAE VC model. The encoder and decoder network utilize the multi-scale Convolutional Neural Network (CNN) architecture [13] as shown in Fig. 2. In addition to the basic VAE framework, the auxiliary gate variable g is introduced to control the amount of the speaker individuality in the output features. The reason for this controlling is that some speech segments, such as silence, may not contain any speaker individuality. By introducing the gating variable, the model can ignore these segments by outputting the gate variable $g = 0$. The gating variable is inferred directly on the input features by the *individuality detector* block.

IV. EXPERIMENTS

A. Dataset

We used the VCTK corpus [14], 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 consisted 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 speaker from the training set (seen speakers).

As speech features, we used WORLD vocoder [16] to extract F_0, spectral sequence, and aperiodicity from speech waveform. Then the spectral sequence is transformed to 60^{th}-order Mel-cepstral coefficients (MCC). Since the spectral envelope from WORLD vocoder is very smooth, high-order cepstral coefficients, which capture the fine fluctuation in the spectral envelope, can be neglected during the conversion process. Therefore, we used only the 1^{st} to 31^{th} MCC coefficients along with interpolated F_0 and voice/unvoiced flag as the input features.

For the proposed system, 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.

For the baseline models, the GMM-based VC (denoted as GMM) used in Voice Conversion Challenge 2018 and the VAE-based VC model that uses the fixed one-hot encoded speaker vector (OH-VAE) [10] is used. For the baseline onehot-VAE model, we keep most of the model architecture identical to the proposed model for a fair comparison. Since the baseline model cannot convert voice to unseen target speaker, we only evaluate the baseline model in seen source to seen target (*seen-seen*) and unseen source to seen target

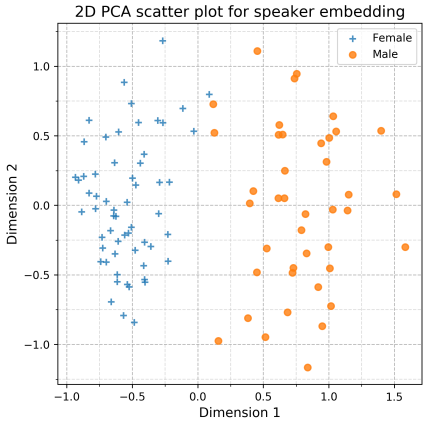

Fig. 3. Learned speaker embedding map of VCTK dataset

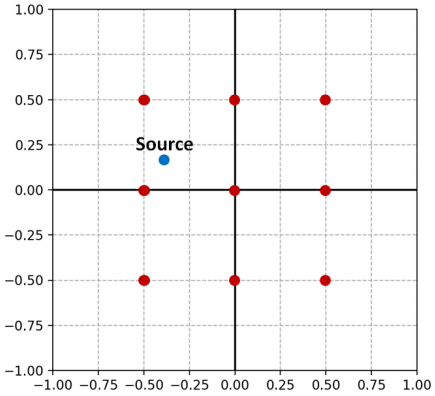

Fig. 4. **Blue**: Position of source speaker embedding vector, **Red**: Position of selected target speaker embedding vector for synthesizing fictitious voices

(*unseen-seen*) conversion scenarios. To perform voice conversion, the baseline VAE model uses the one-hot encoded vector, while the proposed model uses the speaker embedding extracted from a 10-second utterance of the target speaker. For each source-target speaker pair, we trained a separate GMM-based VC model using 100 pair of parallel utterances.

B. Speaker Embedding Space

After the VC model is trained, we visualize the speaker embedding space by analyzing the speaker embedding using Principle Component Analysis (PCA). 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.

C. Fictitious Speaker

We input the speaker embedding vector that is sampled from the speaker embedding space to obtain the fictitious voices that are not available 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) with the position on speaker embedding space shown in Fig. 4.

TABLE I
RESULT OF MOS TEST FOR SPEECH NATURALNESS IN INTRA-GENDER AND CROSS-GENDER CONVERSION

	GMM	OH-VAE	SE-VAE (*proposed*)
F-F	1.55±0.32*	3.06±0.67	**3.14±0.40**
M-M	1.66±0.43*	3.31±0.71	**3.30±0.74**
F-M	1.34±0.24*	2.25±0.35	**2.23±0.42**
M-F	1.48±0.28*	2.88±0.45	**2.72±0.57**
All	1.51±0.16*	2.88±0.32	**2.85±0.32**

D. Subjective Evaluations

To evaluate the quality of the converted waveform, we conducted two listening test: the speech naturalness test and speaker similarity test. Eight listeners (6 males, 2 females) with normal hearing ability enrolled in these tests. All the listeners rate the same sets of test stimuli.

1) Speech naturalness test: We measure the naturalness of converted speech from the baseline models and the proposed model using Mean-Opinion Score evaluation in 5 test scenarios: 1) seen-seen, 2) unseen-seen, 3) seen-unseen, 4) unseen-unseen and 5) fictitious target speaker. Two target speakers (1 male (M), 1 female (F)) were selected for each test scenarios except scenario 5. For each target speakers, all the models will perform both intra-gender (M-M, and F-F) and cross-gender (M-F, and F-M) conversion with the same source speakers. The listeners are instructed to concentrate on the quality of the speech and rate the sample using 5 point-scale that consisted of "bad" (1), "poor" (2), "fair" (3), "good" (4) and "excellent" (5). The order of test stimuli is randomized for each speaker. The result shown in TABLE I and Fig. 5 indicates that the speech waveform generated from the proposed model have higher naturalness than those generated from the GMM-based model in all conversion scenarios. When compared with the onehot-VAE model, the proposed model can synthesize waveform with equivalent naturalness, although only one utterance from the target speaker is required as the reference. The results in TABLE I marked with an asterisk are significantly different ($p < 0.05$) as compared to the proposed SE-VAE model. Moreover, the generated speech of fictitious speakers also has fair naturalness of 3.1 MOS.

2) Speaker similarity test: In this experiment, the speaker similarity between the converted waveform and the target waveform is evaluated in 4 test scenarios: 1) seen-seen, 2) unseen-seen, 3) seen-unseen and 4) unseen-unseen. The listeners are given a reference utterance from target speaker and several converted utterances from different source speakers. All the test stimuli are identical to the test stimuli in naturalness test. The listeners were instructed to concentrate on the voice characteristics and ignore any distortion or degradation in the test stimuli. Then the listener judges the voice similarity between the converted utterances with the reference utterance

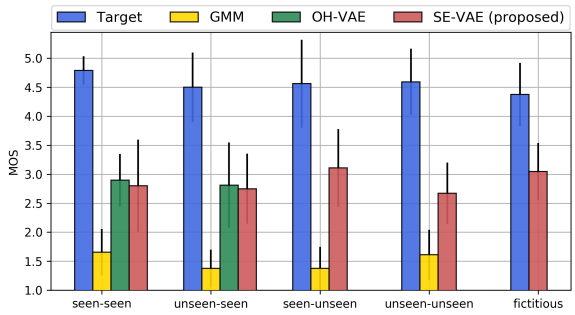

Fig. 5. Result of MOS test for speech naturalness when adapting from seen/unseen-source to seen/unseen-target speaker with 95% confidence interval

TABLE II
RESULT OF MOS TEST FOR SPEAKER SIMILARITY IN INTRA-GENDER AND CROSS-GENDER CONVERSION

	GMM	**OH-VAE**	**SE-VAE** (*proposed*)
F-F	2.07±0.53	2.18±0.62	**2.39±0.65**
M-M	2.11±0.62	2.89±0.98	**2.54±0.72**
F-M	1.77±0.44	2.18±0.69	**1.82±0.72**
M-F	1.52±0.27*	2.93±0.82	**2.57±0.62**
All	1.87±0.25*	2.54±0.41	**2.33±0.35**

using the 5-point scale "not at all similar" (1), "slightly similar" (2), "moderately similar" (3), "very similar" (4) and "extremely similar" (5). The result of the similarity test is reported in TABLE II and Fig. 6, with the asterisk indicates that the different is statistically significant ($p < 0.05$) as compared to the proposed SE-VAE model. From the result, it is clear that there is no difference between the proposed SE-VAE model and OH-VAE, despite the lacks of a large number of training examples from target speaker in the proposed scheme. The overall performance of the proposed VC system is significantly better than the GMM-based VC.

V. CONCLUSIONS

Our proposed method provides a flexible way to control speaker individuality of converted speech by modifying speaker embedding vectors. Although only a single utterance is required as the reference, the subjective test results have shown that the proposed model can convert speech with better perceived naturalness and speaker similarity to the baseline GMM-based model and comparable to the onehot-VAE model. Moreover, since the proposed model can synthesize arbitrary voices with good naturalness, it can be beneficial in various practical application to generate unseen voices. However, we also acknowledge that this technology can be used for the malicious purpose (i.e. voice impersonation) as the proposed model can mimic any voice with only a few seconds of recording. Therefore, countermeasure methods for speaker spoofing attack must also be studied in parallel with this study. The speech samples of this study can be found at [1].

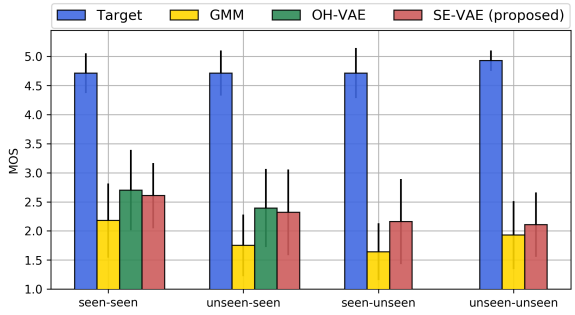

Fig. 6. Result of MOS test for speaker similarity when adapting from seen/unseen-source to seen/unseen-target speaker with 95% confidence interval

ACKNOWLEDGMENT

This study was supported by grants-in-Aid for Scientific Research (B) (No. 17H01761).

REFERENCES

[1] S. H. Mohammadi, A. Kain, "An overview of voice conversion systems," in *Journal of Speech Communication*, vol. 88, pp. 65-82, April 2017.
[2] O. Turk and L. M. Arslan, "Voice conversion methods for vocal tract and pitch contour modification," in *Proceedings of Eurospeech*, pp 2845-2848, 2003.
[3] J. M. G. Arriola, Y. S. Hsiao, J. M. Montero, J. M. Pardo, and D. G. Childers, "Voice conversion based on parameter transformation," in *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, Vol. 3, pp. 987-990, Sydney, Australia, 1998.
[4] M. Akagi and T. Ienaga. "Speaker individuality in fundamental frequency contours and its control," in *Journal of the Acoustical Society of Japan*, vol. 18, no. 2, pp. 73-80, 1997.
[5] T. Toda, A. W. Black, and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory, in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2222-2235, November 2007.
[6] T. Toda, Y. Ohtani, K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. 12491252, April 2007.
[7] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks, *arXiv:1806.02169 [cs.SD]*, June 2018.
[8] 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, in *Proc. APSIPA*, pp. 1-10, 2014.
[9] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. International Conference on Learning Representations (ICLR)*, 2014.
[10] 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*, pp. 1-6, 2016.
[11] A. T. Dinh, A. Kain and K. Tjaden, "Using a Manifold Vocoder for spectral voice and style conversion," in *Proc. of Interspeech*, 2019.
[12] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Modified post-filter to rcover modulation spectrum for HMM-based speech synthesis," in *Proc. of IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 710714, USA, December 2014.
[13] T. V. Ho and M. Akagi, "Speech Accent and Gender Recognition using Dilated Convolution Neural Network with Skip and Residual Connection," in *Proc. of Acoustic Society Japan Spring Meeting*, 2019.
[14] C. Veaux,J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR Voice Cloning Toolkit", 2017, http://dx.doi.org/10.7488/ds/1994.

[1]Demo page: http://www.jaist.ac.jp/ s1820029/apsipa2019

[15] J. L. Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The Voice Conversion Challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Odyssey 2018: The Speaker and Language Recognition Workshop*, pp. 195-202, 2018.

[16] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," in *IEICE Transactions on Information and Systems*, vol. E99, pp. 1877-1884, 2016.