IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 5, 2021, accepted February 8, 2021, date of publication March 2, 2021, date of current version April 1, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3063519

Cross-Lingual Voice Conversion With Controllable Speaker Individuality Using Variational Autoencoder and Star Generative Adversarial Network

TUAN VU HO AND MASATO AKAGI, (Member, IEEE)
Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology (JAIST), Nomi 923-1292, Japan
Corresponding author: Tuan Vu Ho (tuanvu.ho@jaist.ac.jp)

This work was supported in part by the National Institute of Informatics-Center for Robust Intelligence and Social Technology (NII-CRIS), Grant-in-Aid for Scientific Research under Grant 20H04207, and in part by the Japan Society for the Promotion of Science (JSPS)-NSFC Bilateral Joint Research Projects/Seminars under Grant JSJSBP120197416.

ABSTRACT This paper proposes a non-parallel cross-lingual voice conversion (CLVC) model that can mimic voice while continuously controlling speaker individuality on the basis of the variational autoencoder (VAE) and star generative adversarial network (StarGAN). Most studies on CLVC only focused on mimicking a particular speaker voice without being able to arbitrarily modify the speaker individuality. In practice, the ability to generate speaker individuality may be more useful than just mimicking voice. Therefore, the proposed model reliably extracts the speaker embedding from different languages using a VAE. An F0 injection method is also introduced into our model to enhance the F0 modeling in the cross-lingual setting. To avoid the over-smoothing degradation problem of the conventional VAE, the adversarial training scheme of the StarGAN is adopted to improve the training-objective function of the VAE in a CLVC task. Objective and subjective measurements confirm the effectiveness of the proposed model and F0 injection method. Furthermore, speaker-similarity measurement on fictitious voices reveal a strong linear relationship between speaker individuality and interpolated speaker embedding, which indicates that speaker individuality can be controlled with our proposed model.

INDEX TERMS Voice conversion, cross-lingual, controllable speaker individuality, variational autoencoder, generative adversarial network.

I. INTRODUCTION

As a subset of voice transformation, voice conversion (VC) is used to modify the speaker individuality conveyed in speech while keeping the linguistic content unaffected [1]. When the source and target voices are in different languages, a cross-lingual VC (CLVC) model that can efficiently work with multi-lingual input must be used. This type of VC model can be useful in many applications such as personalizing a speech-to-speech translator or language-learning platform. Due to the unavailability of parallel source and target data, a VC model based on conventional mapping methods cannot be used for CLVC. To solve this problem, non-parallel VC models have been actively researched. In contrast to the VC models using conventional mapping approaches, these non-parallel VC models are used to disentangle the linguistic information and speaker individuality from the speech waveform. The source speaker individuality is then swapped with the target one while preserving the linguistic information.

The most straight-forward approach for CLVC is by cascading an automatic speech recognition system and text-to-speech system. As speaker identity and text transcription are both required during the training process, this type of approach can be referred to as a supervised approach. Semi-supervised CLVC can be trained without text transcription by applying regularization on the latent variables representing linguistic content. Therefore, a semi-supervised model can be constructed using inexpensive un-transcribed speech data. Common models for text-independent CLVC are the deep Boltzmann machine [2], [3], autoencoder [4], [5],

The associate editor coordinating the review of this manuscript and approving it for publication was Long Wang.

variational autoencoder (VAE) [6]–[8], and generative adversarial network [9] (GAN). Most CLVC methods only focus on mimicking a target speaker voice without generating new speaker individuality. For certain practical applications, such as customizing audiobook and avatar voices, the ability to actively generate new voice individuality as well as passively mimicking a particular target voice is much more useful than solely mimicking the target voice.

In our previous study [10], we proposed a VAE-based intra-lingual VC model with controllable speaker individuality. By using principal component analysis (PCA), speaker individuality can be derived from speaker embedding. However, this VC model has three drawbacks when applied to a CLVC task. First, the learned speaker-embedding encodes the speaker's language along with other speaker individuality, hence, linguistic information is also affected when modifying the speaker embedding. Second, this model does not model the F0 contour, which can significantly differ between languages. Finally, the training objective of this model does not implicitly guarantee that the output speech carries the desired speaker individuality corresponding to the input speaker embedding. This limitation reduces the speaker similarity between the converted speech and target speech. Moreover, using element-wise mean squared error in the reconstruction loss suggests that the acoustic features follow a normal distribution with no correlation across features. This over-simplified objective often leads to over-smoothing, which results in speech that sounds muffled.

Recently, the StarGAN [11] has been successfully applied for non-parallel multi-speaker VC tasks [12]. The superiority of a GAN over other deep generative models arises from its adversarial training scheme, where a generator and discriminator are simultaneously trained to compete with each other. The training process ends when the generator can generate samples indistinguishable from natural ones. This training scheme avoids the use of mean-squared-error loss, reducing over-smoothing usually found in other VC models. However, the training process of a GAN is often very difficult and unstable, which may degrade converted speech quality. Moreover, the lack of explicit latent modeling in a GAN may discourage the disentanglement between speech content and speaker information, reducing the effectiveness of speaker embedding in controlling the speaker individuality.

Therefore, considering the pros and cons of previous studies, we improved upon our previous VAE-based VC model and designed a model for text-independent CLVC that can both mimic voice and continuously control speaker individuality of generated speech. These improvements are as follows:

- The proposed model uses language embedding to represent the language property of input speech. Therefore, language and speaker-individuality factors can be disentangled.
- The value of F0 in logarithmic frequency scale ($logF0$) is directly injected into the decoder to enhance the F0 modeling and provide controllability over F0 contour.

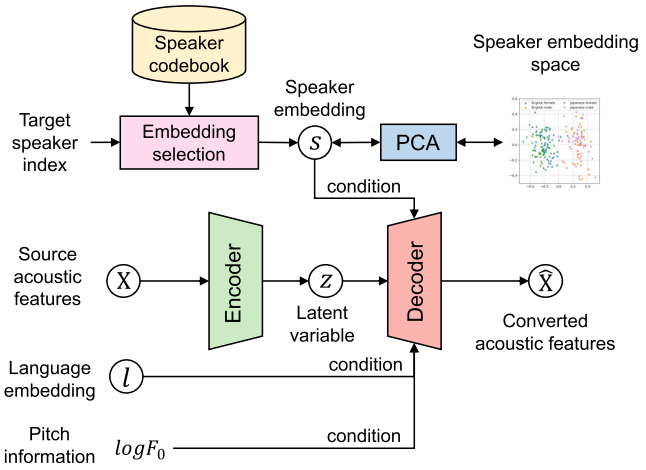

FIGURE 1. Overview of proposed CLVC model. VC is carried out by selecting target speaker embedding from speaker codebook. Each voice characteristic can be independently controlled by PCA-projected speaker embedding.

- The adversarial training scheme of the StarGAN [11] is adopted to improve the objective function of our previous VAE-based VC model.

Although combining the VAE and GAN has been proposed for non-parallel VC [13], [14], none of these studies focused on the controllability of speaker individuality. Our proposed model specifically focuses on the many-to-many CLVC task with controllability of speaker individuality by combining the VAE and StarGAN. To take advantage of the high performance of the recent neural vocoder Parallel WaveGAN [15], our proposed model directly operates in the mel-spectrum domain. Even though continuous speaker embedding has been applied in some VC models [16], [17], they require a trained speaker-recognition model to extract the speaker embedding. In contrast, our proposed model can be trained in an end-to-end fashion by directly optimizing the speaker embedding during the training process. As shown in the next sections, the proposed model improves upon the performance of our previous VAE-based VC model and provides good controllability of speaker individuality by modifying the speaker embedding. Even though our model shares a similar motivation with other VC model regarding F0 conditioning, there are several differences between them. In general, our model focuses on cross-lingual VC settings. As different languages might have very different F0 characteristics, F0 conditioning helps eliminate the language-dependent factor in the speaker embedding. Our previous VAE-based VC mode can still work well without F0 conditioning in an intra-lingual setting [10].

An overview of our proposed model is illustrated in Fig. 1. In Section 2, we discuss related work on VAE- and GAN-based VC models. In Section 3, we describe our proposed CLVC model using the VAE and StarGAN with controllable speaker individuality. We discuss the objective and subjective experiments to evaluate the proposed model and present the results in Section 4. We conclude the paper with a summary in Section 5.

II. LITERATURE REVIEW

In this section, we describe related studies on VAE-based VC models for controlling speaker individuality. Then, we introduce the StarGAN and explain its advantage points that can be adopted to enhance the VAE-based CLVC model.

A. VARIATIONAL-AUTOENCODER-BASED VOICE CONVERSION WITH SPEAKER INDIVIDUALITY CONTROL

1) VARIATIONAL-AUTOENCODER-BASED VOICE CONVERSION

The VAE is a probabilistic model that can discover the latent structure of data [18]. In VC, a previous study by Hsu *et al.* [13] showed that linguistic information can be interpolated via latent representation of the VAE. The latent variable $\mathbf{z}$ is assumed to follow the normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ that is independent from the speaker information. Hence, the latent variable $\mathbf{z}$ can be regarded as linguistic information conveyed in speech. From the input acoustic feature $\mathbf{x}$, the encoder of the VAE f_{enc} outputs the estimated parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ of the posterior $p_\theta(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$. Then $\mathbf{z}$ is sampled from the posterior as $\mathbf{z} \sim p(\mathbf{z}|\mathbf{x})$. However, back-propagation is impossible if $\mathbf{z}$ is directly sampled from the posterior $p_\theta(\mathbf{z}|\mathbf{x})$. Therefore, a re-parameterization trick is applied by sampling an independent variable $\boldsymbol{\epsilon}$ from normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ then executing a scale and shift operation. The procedure of estimating $\mathbf{z}$ is as follows:

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

where $\circ$ is the Hadamard product.

To reconstruct $\mathbf{x}$, in addition to the linguistic information in $\mathbf{z}$, a variable $\mathbf{s}$ that contains speaker information is introduced. The $\mathbf{s}$ can be expressed as a one-hot encoded vector or continuous vector that represents the speaker's identity. From $\mathbf{z}$ and $\mathbf{s}$, the decoder of the VAE reconstructs $\mathbf{x}$s as follows:

$$\widehat{\mathbf{x}} = f_{dec}(\mathbf{z}, \mathbf{s}). \tag{2}$$

The encoder and decoder are jointly trained by minimizing the variational objective function:

$$\mathcal{L}_v = -D_{KL}(p_\theta(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - \mathbb{E}_{\mathbf{z}\sim p_\theta(\mathbf{z}|\mathbf{x})}(p(\mathbf{x}|\mathbf{z}, \mathbf{s})), \tag{3}$$

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

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

The second term on the right side of (3) is the reconstruction loss. Assuming that $\mathbf{x}$ also follows a Gaussian distribution, the term $\mathbb{E}_{\mathbf{z}\sim p_\theta(\mathbf{z}|\mathbf{x})}(p(\mathbf{x}|\mathbf{z}, \mathbf{s}))$ can be described by a simple mean-squared difference between reconstructed acoustic features and original acoustic features as

$$\mathbb{E}_{\mathbf{z}\sim p_\theta(\mathbf{z}|\mathbf{x})}(p(\mathbf{x}|\mathbf{z}, \mathbf{s})) = -\frac{1}{2}\sum(\widehat{\mathbf{x}} - \mathbf{x})^2. \tag{5}$$

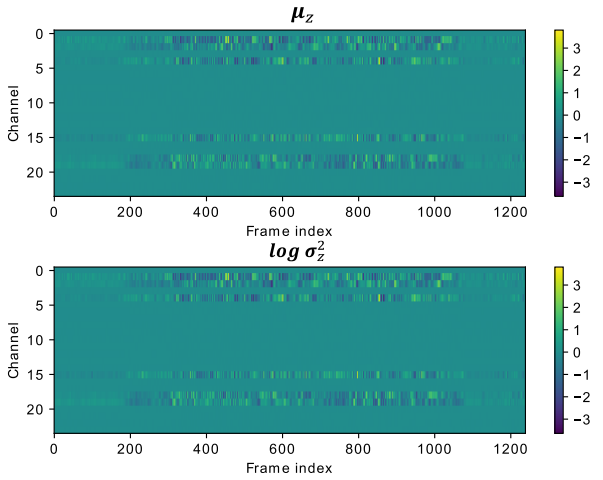

FIGURE 2. Generated parameters for posterior $q(\mathbf{z}|\mathbf{x})$. Most dimensions of latent mean $\mu_{\mathbf{z}}$ and log variance $\log\sigma^2$ are invariant with respect to input.

According to Rolinek *et al.* [19], the optimization of (3) will lead to a *polarized regime* situation, in which only a subset of the latent variables (active subset) encodes meaningful information, while the other subset (passive subset) purely encodes noise. Clearly, the passive subset has $D_{KL} \approx 0$. Therefore, the second term in (3) encourages a bottleneck in the latent variable, where useful information is restricted only in the active subset. Figure 2 illustrates the inferred latent statistical parameters from an input utterance. Since most of the dimensions are invariant with $\mathbf{x}$, the decoder is unable to fully reconstruct the $\mathbf{x}$s without any additional information. In this situation, the decoder network has to rely on the speaker information contained in the input speaker embedding to minimize the reconstruction loss (second term in (3)). This is the cause of the disentanglement of linguistic information and speaker information in the VAE.

2) CONTROLLING SPEAKER INDIVIDUALITY

In a previous study [20], the speaker identity $\mathbf{s}$ was represented as a one-hot vector. However, this representation cannot be used to continuously control the degree of speaker individuality. To solve this problem, we previously proposed a continuous speaker embedding that can be optimized simultaneously with other model parameters [10]. Let $\mathbf{y}$ be the one-hot vector representing speaker identity, the continuous speaker embedding $\mathbf{s}$ is calculated using a simple linear transformation as

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

where $\mathbf{W}$ and $\mathbf{b}$ is a learnable kernel and bias in a fully-connected neural network layer, respectively. In this interpretation, the one-hot encoded vector $\mathbf{y}$ acts as a switch to select the corresponding row vector in matrix $\mathbf{W}$. In the case of $\mathbf{b} = \mathbf{0}$, each row vector in the kernel matrix $\mathbf{W}$ can be seen as a speaker embedding. Figure 3 illustrates the first and second principal components of the learned speaker embeddings of the Voice Cloning Toolkit (VCTK) dataset [21].

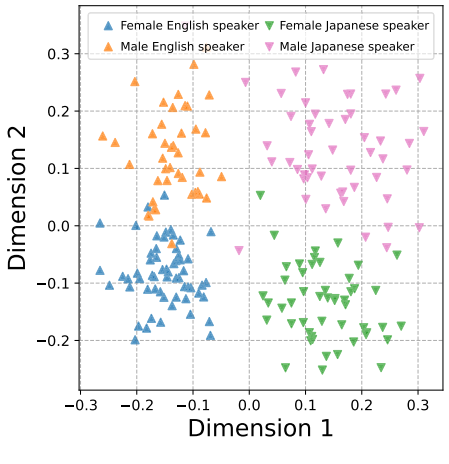

FIGURE 3. 2D visualization of speaker embedding learned using intra-lingual VAE-based VC model using PCA. speaker embeddings are clustered on basis of voice gender and speaker language.

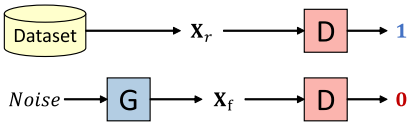

FIGURE 4. Example of GAN consisting of generator (*G*) and discriminator (*D*). D distinguishes real sample $\mathbf{X}_r$ and fake sample $\mathbf{X}_f$, which is generated from *G*. In contrast, *G* generates more realistic fake sample that can deceive *D*.

The speakers are clearly clustered on the basis of the voice gender and input language; hence, the speaker embedding can encode useful information about speaker individuality. However, the language-dependent speaker embedding is not ideal for CLVC as modifying the speaker embedding might affect the linguistic content of the input speech (e.g., unnatural pronunciation).

B. STAR-GENERATIVE-ADVERSARIAL-NETWORK-BASED VOICE CONVERSION

A typical GAN consists of two networks, a generator G and discriminator D, which are alternatively trained to compete with each other in an adversarial scheme [9]. On one hand, D is trained to distinguish between the real sample from the training set and the fake sample from G. On the other hand, G is trained to generate samples that could deceive D. Figure 4 presents an overview of the conventional GAN structure. The model is converged when D exceeds its capability of classifying the generated samples from real samples. In such a situation, G is expected to generate highly realistic samples.

The conventional GAN can only convert data from one domain to another. To solve the problem of multi-domain generation, the StarGAN [11] was proposed. The goal with the StarGAN is to learn a single G that can map across multiple domains. To achieve this, G is trained to translate the input speech features $\mathbf{x}_r$ into output speech features $\mathbf{x}_f$ conditioned on the target domain label $\mathbf{y}_f$, such that $G(\mathbf{x}_r, \mathbf{x}_f) \rightarrow \mathbf{x}_f$. The target domain label is randomly generated to ensure that G can flexibly translate the input data to different target

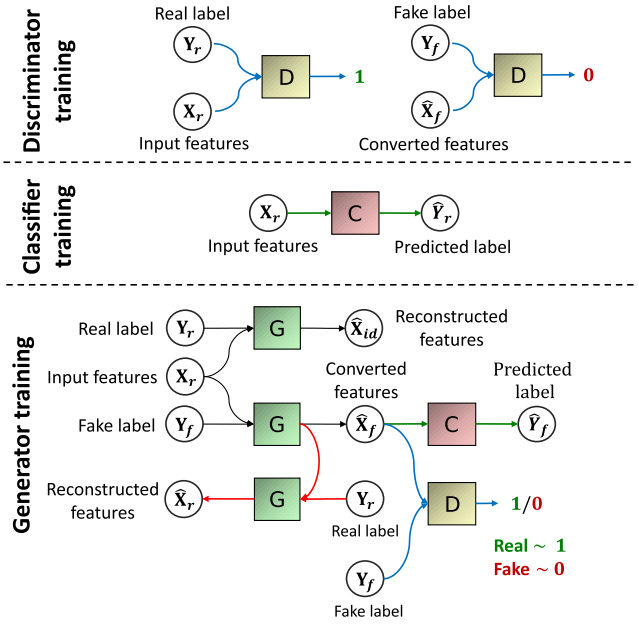

FIGURE 5. Flow chart of StarGAN training process.

domains. Simultaneously, D is trained to estimate the probability $D(\mathbf{x}, \mathbf{y})$ of whether $\mathbf{x}$ is authentic, conditioned on $\mathbf{y}$ of the input data. Also, an auxiliary classifier C is trained to predict this label. Figure 5 shows the training process of the StarGAN. The training objective consists of three loss functions, as detailed below.

- **Adversarial loss**: Adversarial loss encourages D to correctly classify real and fake samples while helping G to generate more realistic samples. The adversarial losses for D and G are respectively as follows:

$$\begin{aligned}\mathcal{L}_{adv}^{D} &= -\mathbb{E}_{\mathbf{x}_r,\mathbf{y}_r}[\log D(\mathbf{x}_r, \mathbf{y}_r)] \\ &\quad - \mathbb{E}_{\mathbf{x}_r,\mathbf{y}_f}[\log(1 - D(G(\mathbf{x}_r, \mathbf{y}_f)), \mathbf{y}_f))], \quad (7)\\ \mathcal{L}_{adv}^{G} &= -\mathbb{E}_{\mathbf{x}_r,\mathbf{y}_f}[\log(D(G(\mathbf{x}_r, \mathbf{y}_f), \mathbf{y}_f). \quad (8)\end{aligned}$$

  The $\mathcal{L}_{adv}^{D}$ is reduced when D can correctly classify real and fake samples, while $\mathcal{L}_{adv}^{G}$ is minimized when G can successfully deceive G.
- **Classification loss**: The C is trained for the speaker-classification task and helps G produce fake data with the correct target speaker voice. In particular, C outputs the probability p_C that $\mathbf{x}$ belong to speaker y. The losses for C and G are defined as

$$\begin{aligned}\mathcal{L}_{cls}^{C} &= -\mathbb{E}_{\mathbf{x}_r,\mathbf{y}_r}[\log p_C(\mathbf{y}_r|\mathbf{x})], \quad (9)\\ \mathcal{L}_{cls}^{G} &= -\mathbb{E}_{\mathbf{x}_r,\mathbf{y}_f}[\log p_C(\mathbf{y}_f|G(\mathbf{x}_r, \mathbf{y}_f))]. \quad (10)\end{aligned}$$

  The $\mathcal{L}_{cls}^{C}$ is reduced when C can correctly classify to which target speaker the input speech belongs. The $\mathcal{L}_{cls}^{G}$ is minimized when the converted utterance has similar speaker individuality to the target speaker.
- **Reconstruction loss**: To preserve the linguistic content in the converted utterance, cycle-consistent loss is introduced to regularize G:

$$\mathcal{L}_{cyc}^{G} = \mathbb{E}_{\mathbf{x}_r,\mathbf{y}_r,\mathbf{y}_f}[||\mathbf{x}_r - G(G(\mathbf{x}_r, \mathbf{y}_f), \mathbf{y}_r)||_2^2], \quad (11)$$

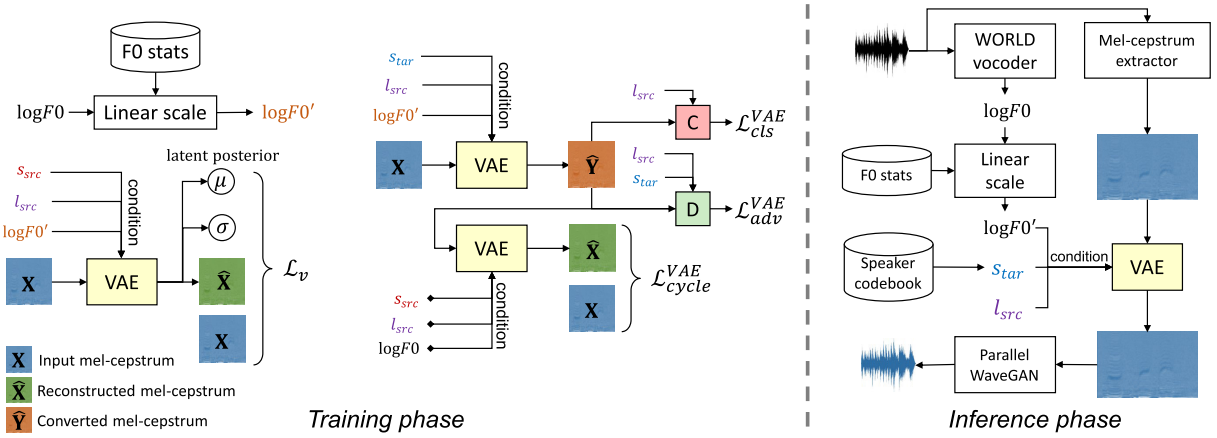

FIGURE 6. Overview of processing flow of proposed model. VAE acts as *G* of StarGAN. l_{src} refers to language embedding of input mel-cepstrum. s_{src} and s_{tar} are speaker embedding of source and target speakers.

where $\mathbf{y}_r$ and $\mathbf{y}_f$ are the labels of arbitrary source and target speaker, respectively, $\mathbf{x}_r$ is the input speech feature belonging to $\mathbf{y}_r$, and $\|\cdot\|$ is the Euclidean distance.

Identity loss is also introduced to keep the converted speech unchanged when the input speech already belongs to $\mathbf{y}_r$:

$$\mathcal{L}_{id}^{G} = \mathbb{E}_{\mathbf{x}_r,\mathbf{y}_r}[||\mathbf{x}_r - G(\mathbf{x}_r, \mathbf{y}_r)||_2^2]. \tag{12}$$

In summary, the total loss for G is as follows:

$$\mathcal{L}^{G} = \mathcal{L}_{id}^{G} + \mathcal{L}_{cyc}^{G} + \lambda_{adv}\mathcal{L}_{adv}^{G} + \lambda_{cls}\mathcal{L}_{cls}^{G}, \tag{13}$$

where λ_{adv} and λ_{cls} are the weighting factor for adversarial loss and classifier loss, respectively.

As seen in the training objective (13), the StarGAN does not completely rely on mean-squared-error loss to estimate the distribution of converted acoustic features, as in the VAE. In contrast, G uses feedback from D to produce the most likely sample that can deceive D. Therefore, to avoid over-smoothing in the VAE, the adversarial training scheme of the StarGAN can be adopted to replace the conventional mean-squared-error loss. However, the lack of an explicitly defined latent variable in the StarGAN might reduce the effect of speaker embedding on controlling speaker individuality because G might ignore the input speaker embedding. Hence, the combination of the VAE and StarGAN would alleviate the weakness of the other.

III. PROPOSED MODEL

In this section, we give a more detailed explanation of the proposed CLVC model. We first describe the solution to avoid the language-dependent speaker-embedding problem of our previous VAE-based VC model. We then present the F0 injection method to enhance the F0 modeling in different languages. Finally, we introduce a method for enhancing the spectral detail using the StarGAN.

A. CONTROLLING SPEAKER INDIVIDUALITY IN CROSS-LINGUAL SETTING

In conventional VAE-based VC, speaker identity is usually represented as a one-hot vector [20]. However, this type of encoding does not allow controllability of speaker individuality. Some studies have proposed using d-vector to represent speaker individuality, but this type of speaker representation requires an additional speaker-recognition network, which introduces more complexity to the VC model. Our previous VAE-based VC model was developed for continuous learnable speaker embedding that can be jointly learned with other network parameters during the training process [10]. This model does not require any addition speaker-recognition network yet still achieves controllability of speaker individuality.

In this study, we improved upon this model for cross-lingual settings by training the VAE on cross-lingual data. We simplify (6) by setting $\mathbf{b} = \mathbf{0}$; hence, the kernel $\mathbf{W}$ can be regarded as the speaker codebook, which is randomly initialized. During inference, only speaker embedding of the target speaker is needed for conditioning the decoder network. However, language differences can be captured during the speaker embedding, as shown in Fig. 3. This behavior is undesirable because manipulating the speaker embedding would affect the linguistic content due to language differences. To avoid this problem, we implement an additional language embedding to disentangle the language factor from speaker embedding. In this study, the language factor is simply represented by a one-hot encoded vector, which is concatenated with the speaker embedding along the channel dimension. The combined vector is then used to condition the decoder on generating the mel-spectrogram, as shown in Fig. 6.

B. ENHANCING F0 MODELING WITH F0 INJECTION

Various high-performance vocoders based on deep neural networks have recently been proposed [15], [22], [23]. Most

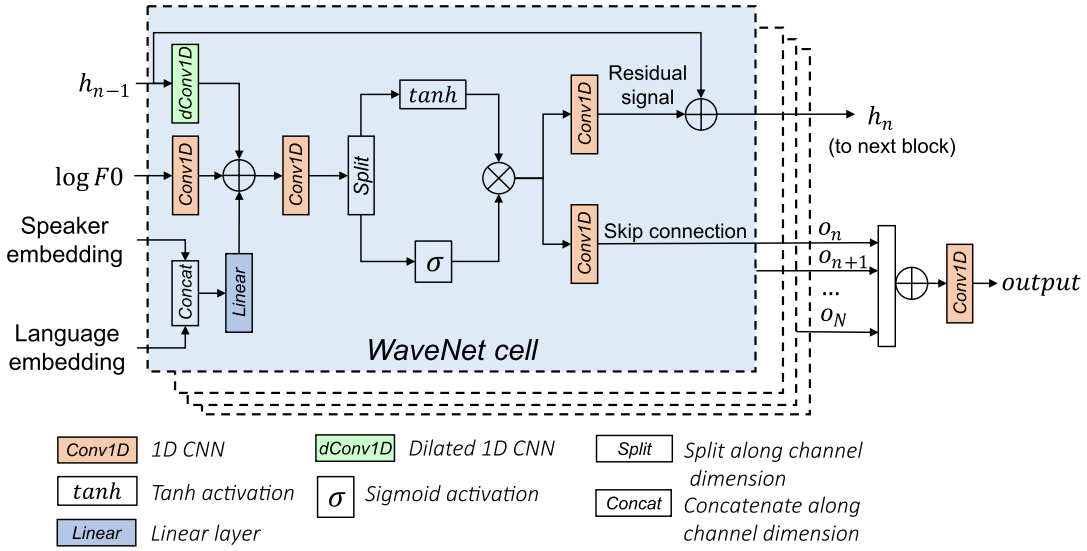

FIGURE 7. Stacks of WaveNet cells in WaveNet module.

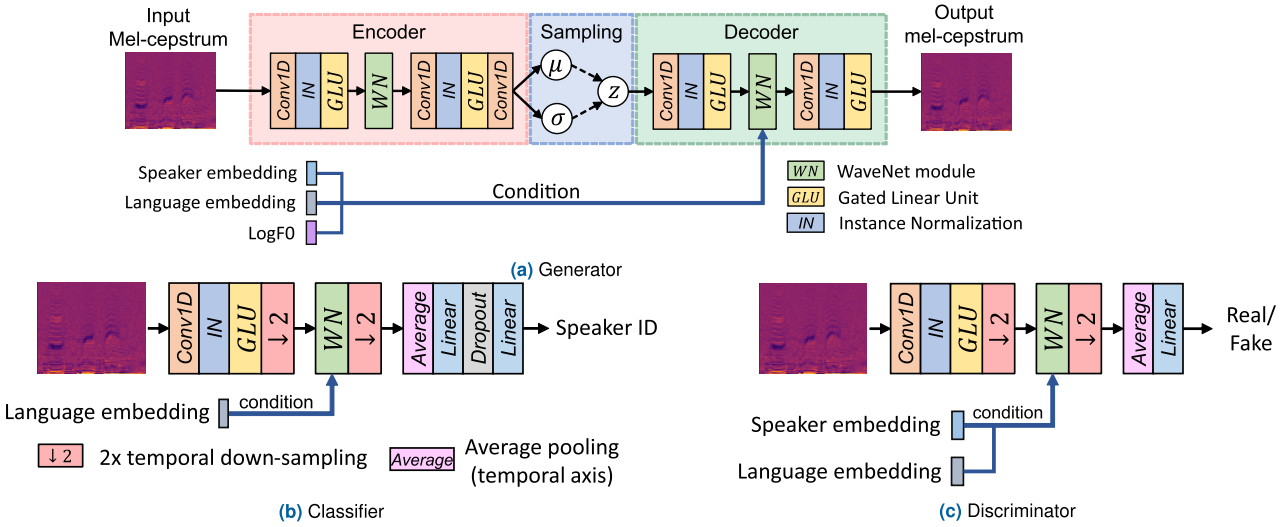

FIGURE 8. Structure of (a) *G*, (b) classifier (*C*), and (c) *D*.

of these neural vocoders directly use mel-spectrum as the input feature. However, it is difficult to directly manipulate the F0 information in mel-spectrogram, as it relates to the harmonic structure. In addition, different languages may have very different F0 contours, which can degrade the cross-lingual converted speech with spurious pitch. To provide the controllability and stability of F0 in converted speech, we directly conditioned the decoder in the VAE with $\log F0$ input, as shown in Figs. 6 and 7. We refer to this method as *F0 injection*. To generate fake samples during the training or inference phases, the source $\log F0$ is linearly scaled to match the target $F0$ mean-variance. Therefore, the statistics of the target $F0$ must be pre-calculated for VC.

C. IMPROVING CROSS-LINGUAL VAE-BASED VC WITH StarGAN TRAINING SCHEME

Our proposed model incorporates the StarGAN training scheme [11]. An overview of our proposed model is shown in Fig. 6. In this model, the VAE acts similarly to the *G* in the StarGAN. The *D* identifies whether the input speech is natural or converted given the speaker-identity label. The *C* learns to classify to which speaker the input speech belongs. Also, the converted voice is re-input to the VAE to convert it back to the source voice. Cycle-consistent loss minimizes the difference between the input features and re-converted features. With all these modifications, the new training objective for the VAE is to 1) generate converted speech to deceive *D*, 2) minimize the loss from *C* when inputting the converted

speech, 3) minimize cycle-consistent loss, and 4) minimize reconstruction loss and D_{KL} loss.

- **Discriminator loss**: The D distinguishes real and converted speech samples, which are labeled as $\mathbf{1}$ and $-\mathbf{1}$, respectively. To improve the stability of the training process, the Wasserstein distance [24] is used instead of vanilla discriminator loss in (7). Therefore, discriminator loss is written as

$$\mathcal{L}_{adv}^{D} = \mathbb{E}_{\mathbf{x},\mathbf{s}_{src}}[1 - D(\mathbf{x}, \mathbf{s}_{src})] + \mathbb{E}_{\mathbf{x},\mathbf{s}_{tar}}[1 + D(\mathrm{VAE}(\mathbf{x}, \mathbf{s}_{tar}), \mathbf{s}_{tar})], \quad (14)$$

  where $\mathbf{s}_{src}$ and $\mathbf{s}_{tar}$ is the speaker embedding of source and target speakers, respectively, and x is the input acoustic features belonging to the source speaker.
- **Classification loss** The C is trained with cross-entropy loss to identify the correct speaker identity conveyed in the input utterance. The loss for training C is as follows:

$$\mathcal{L}_{cls}^{C} = -\mathbb{E}_{\mathbf{x},\mathbf{y}}[\log p_C(\mathbf{y}|\mathbf{x})] \quad (15)$$

  where $\log p_C(\mathbf{y}|\mathbf{x})$ is the output log likelihood that acoustic features x belongs to target speaker $\mathbf{y}$.
- **VAE loss**: In addition to variational loss, adversarial loss and classifier loss encourage the VAE to trick D and reduce the speaker dissimilarity between converted speech and natural speech. The adversarial loss and classifier loss for the VAE are expressed as

$$\mathcal{L}_{adv}^{\mathrm{VAE}} = -\mathbb{E}_{\mathbf{x},\mathbf{s}_{tar}}[D(\mathrm{VAE}(\mathbf{x}, \mathbf{s}_{tar}), \mathbf{s}_{tar})], \quad (16)$$

$$\mathcal{L}_{cls}^{\mathrm{VAE}} = -\mathbb{E}_{\mathbf{x},\mathbf{s}_{tar}}[\log p_C(\mathrm{VAE}(\mathbf{x}, \mathbf{s}_{tar}), \mathbf{s}_{tar})]. \quad (17)$$

  Similar to the StarGAN training scheme, cycle-consistent loss is introduced to force the VAE to transform the converted features back to the original. This loss is written as

$$\mathcal{L}_{cycle}^{\mathrm{VAE}} = \mathbb{E}_{\mathbf{x},\mathbf{s}_{tar},\mathbf{s}_{src}}[\|\mathbf{x} - \mathrm{VAE}((\mathrm{VAE}(\mathbf{x}, \mathbf{s}_{tar}), \mathbf{s}_{src})\|_2^2]. \quad (18)$$

  Combined with the variational loss described in (3), the final training objective for the proposed model now becomes

$$\mathcal{L}_{obj}^{\mathrm{VAE}} = \mathcal{L}_{v} + \mathcal{L}_{cycle}^{\mathrm{VAE}} + \lambda_{adv}\mathcal{L}_{adv}^{\mathrm{VAE}} + \lambda_{cls}\mathcal{L}_{cls}^{\mathrm{VAE}}, \quad (19)$$

  where λ_{adv} and λ_{cls} are the weight factor for each loss component. In empirical testing, $\lambda_{adv} = 0.0005$ and $\lambda_{cls} = 0.0001$ showed good results in this study.

IV. EXPERIMENTS

To evaluate the performance of the proposed model, we implemented CLVC between English and Japanese speakers using three models: the conventional VAE (**VAE**), StarGAN (**StarGAN**), and proposed model (**VAE-StarGAN**). To evaluate the effectiveness of F0 injection, we also implemented a VAE-based VC model trained without F0 input. This model is denoted as **VAE-noF0**. For a fair comparison, **VAE** and **VAE-StarGAN** had the same network structure. In addition, the C and C of **StarGAN** and **VAE-StarGAN** had an identical structure.

TABLE 1. Network architecture of VAE encoder and decoder, *D*, and *C*.

Network	No. of WN cells	Dilation rate	Filters	Kernel size
Encoder	6	2×[1, 2, 4]	128	5
Decoder	16	4×[1, 2, 4, 8]	128	5
D	3	[1, 1, 1]	[128, 256, 512]	3
C	3	[1, 1, 1]	[128, 256, 512]	3

To train the models, we used two open-source multi-speaker voice databases: the English VCTK corpus [21] and the Japanese Versatile Speech (JVS) corpus [25]. The training data included 100 speakers from the English VCTK dataset and 100 speakers from the JVS dataset. For each speaker, 100 utterances were randomly selected as training data and ten utterances as testing data. Each speaker was initially assigned to a random speaker embedding. To condition the decoder on the language of the input mel-cepstrum, we used a one-hot embedding vector for language. Since there were two input languages (English and Japanese), the number of dimensions for language embedding was two.

A. PREPROCESSING

In the preprocessing step, the audio waveform was downsampled to 24 kHz and normalized to the $[-1.0, 1.0]$ range. Then, an 80-dimensional mel-spectrogram was extracted using short-time Fourier transform (STFT) and mel-filterbank. The window length of STFT was set to 2048 and the hop-length was 300. The mel-filterbank spanned from 80 to 7600 Hz to match the Parallel WaveGAN input. Then, the mel-spectrum was transformed into mel-cepstrum by applying inverse discrete Fourier transform on the log-magnitude mel-spectrum. Although some studies further normalized each mel channel by its mean and variance across the time dimension, we found that this step degrades the quality of converted speech from our models. Therefore, we directly used the raw mel-cepstrum value as the input feature. In addition to the mel-cepstrum feature, $F0$ was extracted using the WORLD analysis system [26]. After extracting the F0 from all utterances, we calculated the mean and variance of $\log F0$ for each speaker for linear scaling functions. To reconstruct the waveform, we used the Parallel WaveGAN vocoder [15] trained on the VCTK dataset for 1000k iterations.

B. NETWORK ARCHITECTURE

Similar to our previous study [10], the encoder and decoder of the VAE were constructed from a smaller network that resembles the WaveNet (WN) architecture [27]. Figure 7 shows the architecture of a WN cell. The input layer for the hidden variable h_n is the 1D dilated convolutional neural network [28], which expands the receptive field in the temporal dimension by dilation in the kernel. The details of the model parameters of the VAE encoder and decoder, D, and C are provided in Table 1.

The D and C share the same architecture, as illustrated in Fig. 8. Each WN cell is followed by a stride 1D convolution layer to reduce the temporal dimension by half after each stage. At the output, a fully connected layer consumes the compressed vector to produce the output vector. The speaker embedding and language embedding are represented as a one-hot vector. Both D and C are conditioned on both the speaker-embedding and language-embedding vectors, while C is conditioned only on the language embedding vector.

C. TRAINING PROCEDURE

All models were trained using the Adam optimizer [29] with 32 samples per batch. The mel-cepstrum is truncated or warp-padded to have 512 frames. The learning rate is initialized at 2×10^{-4} and gradually reduced to 1×10^{-4} for the first ten epochs. The training process was conducted using two Nvidia 2080Ti GPUs until the model converged, which took roughly two days for each model. The detailed training procedure for **StarGAN** and **VAE-StarGAN** is shown in Algorithm 1.

Algorithm 1 VAE-StarGAN training procedure

Require: Functions G (VAE model), D, C, $Scale$ ($\log F0$ linear scale function), $\mathbf{X}$ (batch of source mel-cepstrum), $f0$ (batch of source $\log F0$), $\mathbf{s}_{src}$ (source speaker embedding), $\mathbf{s}_{tar}$ (target speaker embedding), $\mathbf{l}_{src}$ (source language embedding)

▷ Update the discriminator parameter θ_D

$f0_f \leftarrow Scale(f0)$

$\mathbf{x}_f \leftarrow G(\mathbf{x}, F0_f, \mathbf{s}_{tar}, \mathbf{l}_{src})$

$\mathbf{d}_r \leftarrow max(0, 1 - D(\mathbf{x}, \mathbf{s}_{src}, \mathbf{l}_{src}))$

$\mathbf{d}_f \leftarrow max(0, 1 + D(\mathbf{x}_f, \mathbf{s}_{tar}, \mathbf{l}_{src}))$

$\mathcal{L}^D_{adv} \leftarrow \frac{\mathbf{d}_r + \mathbf{d}_f}{2}$

update θ_D to minimize $\mathcal{L}^D_{adv}$

▷ Update classifier parameter θ_C

$\mathcal{L}^C_{cls} \leftarrow \text{CrossEntropy}(\mathbf{s}_{src}, C(\mathbf{x}, \mathbf{l}_{src}))$ update θ_C to minimize $\mathcal{L}^C_{cls}$

▷ Update VAE parameter θ_{VAE}

$\mathbf{x}_{id}, \boldsymbol{\mu}_z, \boldsymbol{\sigma}_z \leftarrow G(\mathbf{x}, F0, \mathbf{s}_{src}, \mathbf{l}_{src})$

$\mathbf{x}_{cycle} \leftarrow G(\mathbf{x}_f, F0, \mathbf{s}_{src}, \mathbf{l}_{src})$

$\mathcal{L}^{VAE}_{adv} \leftarrow -D(\mathbf{x}_f, \mathbf{s}_{tar}, \mathbf{l}_{src})$

$\mathcal{L}^{VAE}_{cls} \leftarrow \text{CrossEntropy}(\mathbf{s}_{tar}, C(\mathbf{x}_f, \mathbf{l}_{src}))$

$\mathcal{L}^{VAE}_{cycle} \leftarrow \|\mathbf{x}_{cycle} - \mathbf{x}\|_2^2$

$\mathcal{L}_v \leftarrow \|\mathbf{x}_{id} - \mathbf{x}\|_2^2 - \frac{1}{2}(1 + \log \boldsymbol{\sigma}_z^2 - \boldsymbol{\mu}_z^2 - \boldsymbol{\sigma}_z^2)$

▷ Calculate 19

$\mathcal{L}^{\text{VAE}} \leftarrow \mathcal{L}_v + \mathcal{L}^{\text{VAE}}_{cycle} + \lambda_{adv}\mathcal{L}^{\text{VAE}}_{adv} + \lambda_{cls}\mathcal{L}^{\text{VAE}}_{cls}$

update θ_G to minimize $\mathcal{L}^{\text{VAE}}$

D. VISUALIZING SPEAKER EMBEDDING

After the VC model was trained, we visualized the speaker-embedding space, as shown in Fig. 9, by analyzing the speaker codebook using PCA. Figure 9a illustrates the

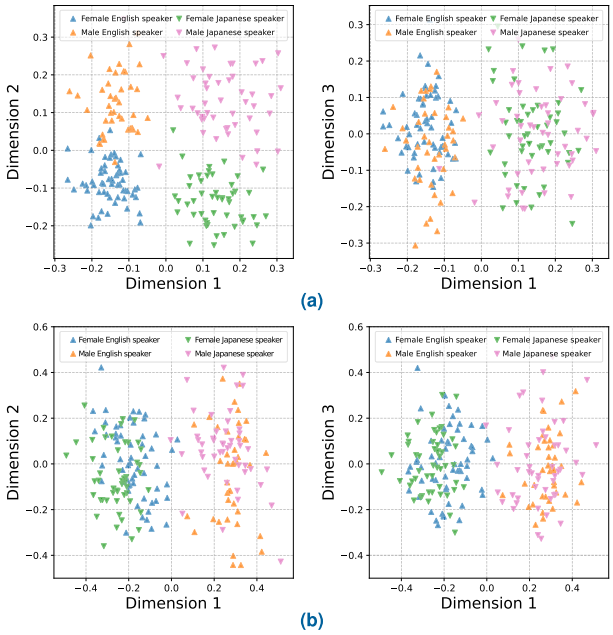

FIGURE 9. 2D PCA visualization of speaker embedding from model (a) without language embedding input and (b) with language embedding input. Speaker embedding from English and Japanese speakers are clearly separated into distinct clusters when language embedding is not used.

PCA-projected speaker embedding learned using our previous VAE-based VC model [10]. Without the input language embedding, we can see that the language of the speakers was separated on the first principal dimension. On the other hand, as shown in Fig. 9b, only the speaker's sex was separated on the first principal dimension when the model was trained with language embedding input. Moreover, the clustering effect on language was removed, as there was no clear separation between Japanese speakers and English speakers. This result indicates that the speaker embedding can encode useful information from the speaker individuality while still remaining language-independent.

E. OBJECTIVE EVALUATION

We conducted different objective measurements to evaluate the performance of the proposed model. The objective evaluation set consists of cross-lingual converted utterances from English to Japanese and Japanese to English. We selected five male and five female speakers from each language to form 200 conversion pairs, and each pair had ten converted samples. Therefore, the objective evaluation set consisted of 2000 converted utterances.

1) MODULATION SPECTRUM MEASUREMENT

The modulation spectrum (MS) can provide hints about speech naturalness: a higher MS corresponds to better speech naturalness. Following the work of Takamichi *et al.* [30], we calculated the MS of the converted mel-cepstral sequence by taking the Fourier transform along the temporal dimension. Similar to a previous study [31], the MS was averaged

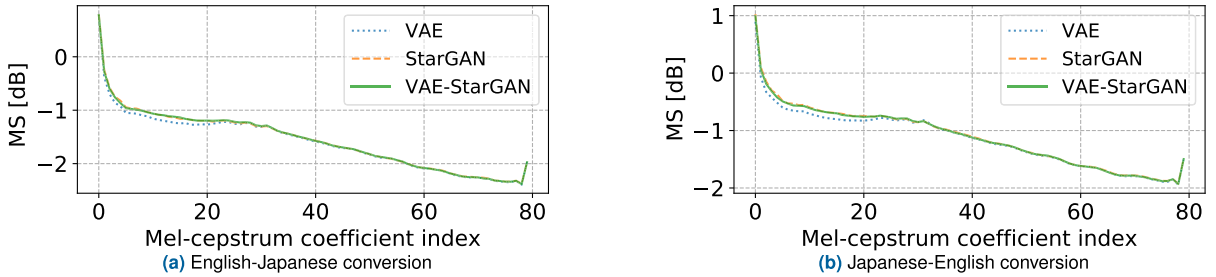

FIGURE 10. Log-scaled modulation spectrum of natural speech, reconstructed speech, and converted speech averaged over all utterances and modulation frequencies. StarGAN and VAE-StarGAN generated mel-spectrograms with higher MS than VAE.

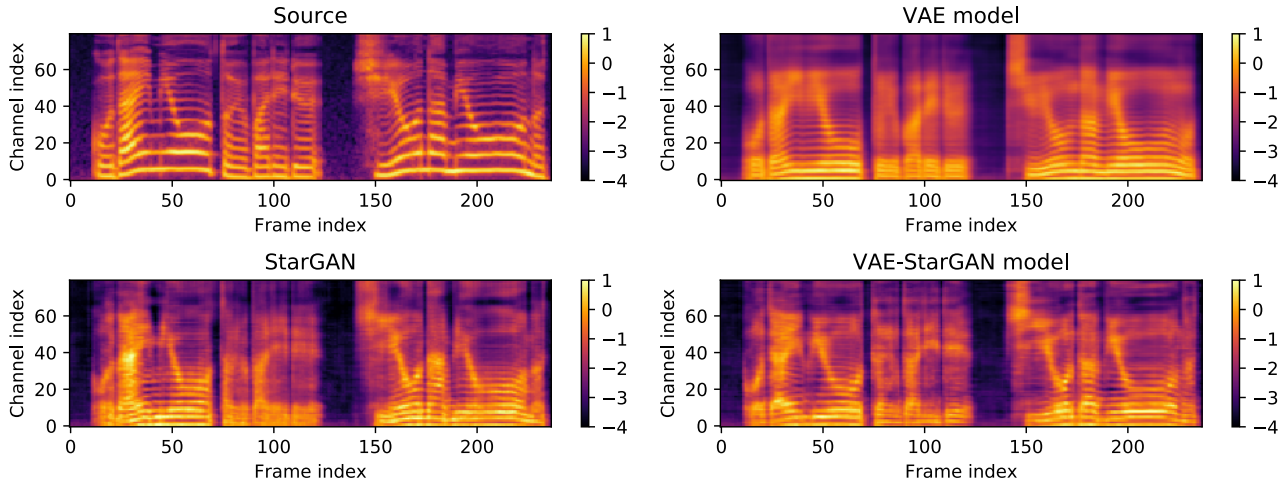

FIGURE 11. Mel-spectrogram of source voice and converted voice from different models. Mel-spectrogram generated from StarGAN and VAE-StarGAN clearly had more details than that generated from VAE.

for all modulation frequencies and all utterances as

$$MS = \frac{1}{N}\frac{1}{F}\sum_{n}^{N}\sum_{f}^{F}|DFT[\mathbf{X}(\mathbf{n},\mathbf{f})]|, \tag{20}$$

where $\mathbf{X}$ is a batch of test utterances, N is the number of utterances, $n \in [0, N)$ is the utterance index, F is the number of MS frequency bins, and $f \in [0, F)$ is the MS frequency bins. As shown in Fig. 10, **VAE-StarGAN** achieved a higher log-scaled MS on the lower mel-cepstral coefficients than our previous VAE-based VC model. These results indicate that the adversarial training scheme can lessen the over-smoothing of converted mel-cepstral coefficients. Figure 11 illustrates the mel-spectrogram generated from different models. We can see that the StarGAN and t VAE-StarGAN produced mel-spectrograms with a more detailed structure. Although the mel-spectrum of **VAE-StarGAN** was more refined than that of VAE, artifacts such as mispronunciation cannot be clearly shown on the mel-spectrum. Therefore, a listening test must be conducted to precisely compare the performances of different models.

2) F0 INJECTION

To measure the effectiveness of F0 injection method, we measured the F0 histogram intersection [32] between converted speech and target speech. The histogram intersection can indicate the amount of similarity between two distributions. Given the histogram of converted speech P and that of target speech Q, where each one contains n bins, the histogram intersection is defined as follows:

$$d_{\cap}(P, Q) = \frac{\sum_{j}^{n} min(P_j, Q_j)}{\sum_{j}^{n} Q_j}. \tag{21}$$

The maximum histogram intersection $d_{\cap max} = 1$ is achieved when P and Q are completely identical. Figure 12 shows a comparison of the $\log_2 F0$ distribution between source, target, and converted utterances from different models. We can see that the $\log_2 F0$ distribution did not always follow the Gaussian shape. Therefore, simply executing F0 linear transformation by a parametric vocoder (e.g., WORLD or STRAIGHT [12], [17], [33]) cannot ensure the correct shape of $F0$ distribution.

In addition to histogram intersection, we measured the average error between the mean of converted $\log_2 F0$ and that of target $\log_2 F0$. The voice/unvoiced error rate between converted F0 and source F0 was also measured. The results are summarized in Table 2. We can see that the models with F0 injection had a significantly higher histogram intersection, lower v/uv error rate, and lower mean F0 error than the model without. The two-tailed t-test showed that the effect of using

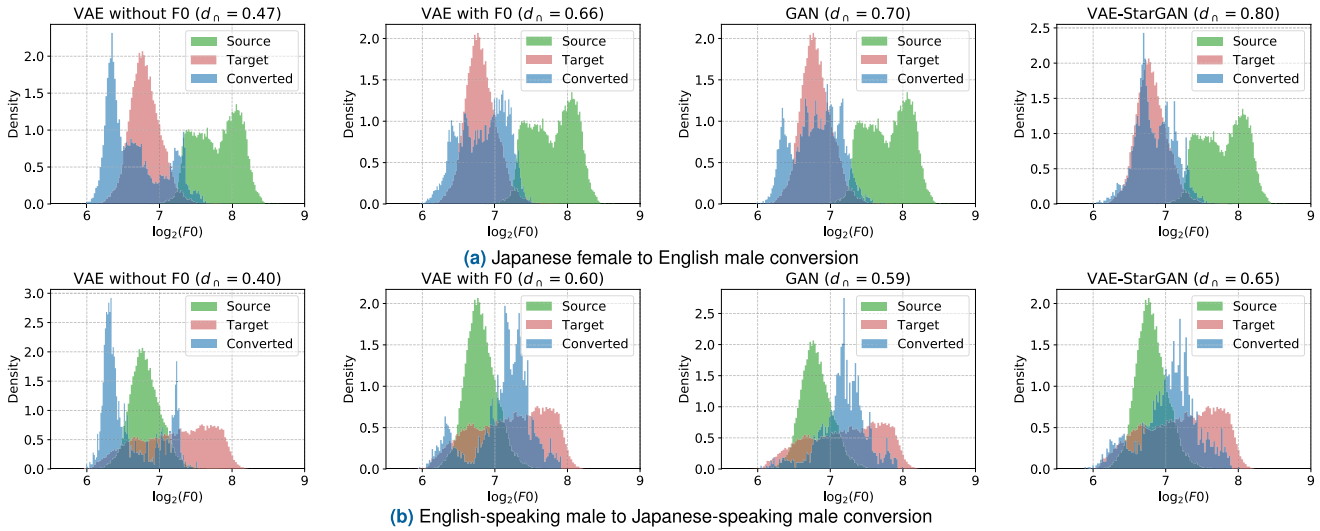

FIGURE 12. Distribution of $\log_2$ *F0* of source, target, and converted speech from different models. Intersection index $d_\cap$ indicates amount of overlap between converted $\log_2$ *F0* and target $\log_2$ *F0*.

TABLE 2. Average scores and standard deviations of *F0* analysis results from different models. For mean F0 error and voice/unvoiced error rate (v/uv) error, lower is better. For histogram intersection, higher is better.

Test / model	All			English to Japanese			Japanese to English		
	Mean F0 error	v/uv error (%)	Histogram intersection	Mean F0 error	v/uv error (%)	Histogram intersection	Mean F0 error	v/uv error (%)	Histogram intersection
VAE-noF0	0.286 ± 0.20	0.191 ± 0.03	0.506 ± 0.08	0.419 ± 0.17	0.198 ± 0.03	0.481 ± 0.07	0.151 ± 0.12	0.184 ± 0.03	0.532 ± 0.08
VAE	0.132 ± 0.09	0.152 ± 0.03	0.563 ± 0.13	0.173 ± 0.07	0.160 ± 0.03	0.564 ± 0.12	0.090 ± 0.09	0.143 ± 0.02	0.562 ± 0.13
StarGAN	$\mathbf{0.082} \pm 0.06$	$\mathbf{0.141} \pm 0.02$	0.568 ± 0.11	$\mathbf{0.080} \pm 0.05$	$\mathbf{0.135} \pm 0.02$	0.567 ± 0.08	0.084 ± 0.07	0.147 ± 0.02	0.568 ± 0.13
VAE-StarGAN	0.128 ± 0.09	0.148 ± 0.02	$\mathbf{0.580} \pm 0.12$	0.173 ± 0.08	0.154 ± 0.02	$\mathbf{0.581} \pm 0.09$	$\mathbf{0.083} \pm 0.08$	$\mathbf{0.143} \pm 0.01$	$\mathbf{0.578} \pm 0.14$

the F0 injection method is statistically significant. These results indicate that the F0 injection method can improve the performance of VC models for controlling the F0 in the converted utterance.

F. SUBJECTIVE EVALUATION

We conducted listening tests to evaluate the speech naturalness and speaker similarity of the converted utterances. We selected one male and one female speaker from each language, for a total of four speakers in the evaluation set. Since only CLVC was carried out, there were eight combinations from the selected speakers. We denote Japanese-to-English conversion as "**SJ-TE**" and English-to-Japanese conversion as "**TE-SJ**". Two sentences were selected from each source-target pair to create the listening test set. Therefore, the listening test set consisted of 48 pairs of converted utterances (2 sentences $\times$ 8 source-target speaker pairs $\times$ 3 model pairs). For reference stimuli in the ABX similarity test, we randomly selected the original utterances of the target speakers from the training set. Nine individuals with normal listening ability participated in both listening tests. All participants had a basic level of using Japanese/English even if Japanese/English was not their first language. Each participant rated 24 random pairs of converted utterances for each test via an online interface.

To measure speaker similarity, the ABX test scheme was used to compare the performance of **VAE-StarGAN**, **StarGAN**, and **VAE**. Listeners were asked to select the closest utterance ("A" or "B") to the reference utterance X or choose *Same* if there was no difference. The X is the natural speech of the target speaker selected from the test set, while utterances "A" and "B" are generated from different models. For speech naturalness, we applied the AB test scheme, in which listeners were asked to determine the more natural utterance ("A" or "B") or choose *Same* if there was no difference. The generated utterance from both models was presented in random order (AB or BA) to avoid any bias. To analyze the results, we used the one-way ANOVA test with alpha value of 0.05.

As shown in Figs. 13 and 14, **VAE-StarGAN** outperformed **StarGAN** for both naturalness and similarity in all cases. Except for the similarity score of SE-TJ conversion, these differences are statistical significant. When comparing with the **VAE**, the one-way ANOVA test and the post-hoc two-tailed t-test determined that **VAE-StarGAN** had a statistically better similarity score than **VAE** in SJ-TE conversion. However, no significant difference was observed between these two models in other cases. **VAE** had better naturalness and similarity scores than **StarGAN** in most cases except for the SE-TJ similarity score. The reason might be that

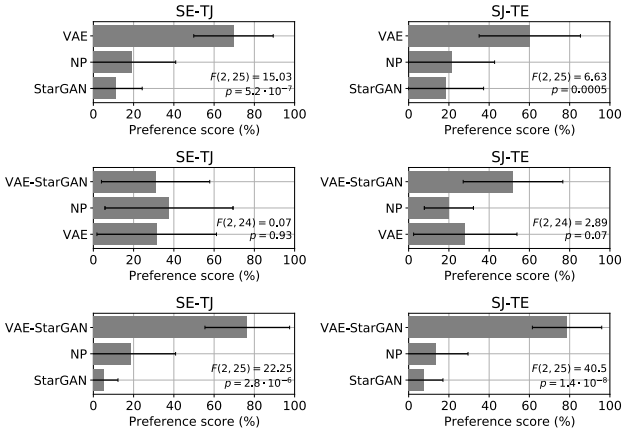

FIGURE 13. Preference scores of AB naturalness test with 95-percent confidence interval and results from one-way ANOVA test. NP means no preference.

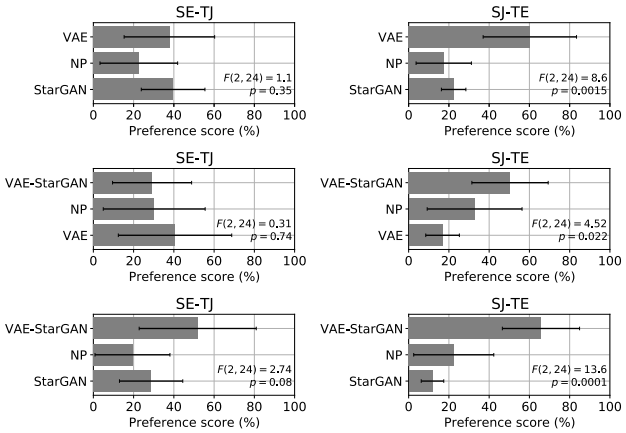

FIGURE 14. Preference scores of ABX speaker similarity test with 95-percent confidence interval and results from one-way ANOVA test.

although the converted speech from **StarGAN** sounded less muffled than that from **VAE**, artifacts such as mispronunciation severely affected the perceived speech naturalness. The low preference score of **VAE-StarGAN** for speaker similarity indicates that the speaker embedding of **StarGAN** has less controllability on speaker individuality than **VAE** and **VAE-StarGAN**. This behavior may be due to the lack of explicit latent modeling in **StarGAN**, which discourages the disentanglement between speech content and speaker information.

G. FICTITIOUS SPEAKER

To evaluate the controllability of speaker individuality with **VAE-StarGAN**, 11 converted utterances were generated by linearly interpolating the speaker embedding between the source and target speaker embeddings. The source speaker was a female Japanese speaker and the target speaker was a male English speaker. The positions of the interpolated speaker embedding **s** are shown in Fig. 15. The input F0 was also transformed using the linearly interpolated mean and standard deviation between the source and target F0.

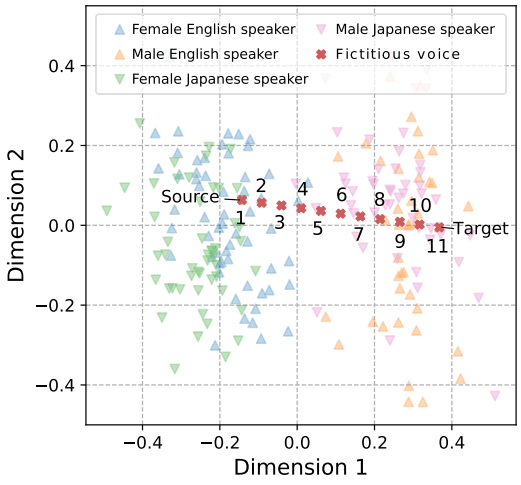

FIGURE 15. Position of linearly interpolated speaker embedding between source female Japanese speaker and target male English speaker. Index of each converted utterance is marked from 1 to 11.

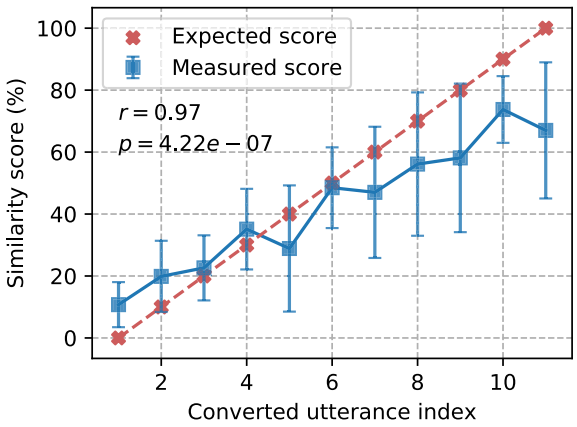

FIGURE 16. Similarity scores of interpolated speaker embedding with standard deviation. Dotted line denotes expected similarity score that linearly increased from 0 to 100. *r* and *p* indicate Pearson correlation and p-value, respectively.

Each test utterance was marked from 1 to 11 with respect to its position on the speaker-embedding map. In this test, the participants listened to the test stimuli in random order to avoid any bias then were asked to judge the similarity between test stimuli and the reference utterance on a scale from 0 to 100. Figure 16 shows the average similarity score of each test utterance. We used the Pearson correlation coefficient to evaluate the linear relationship between average similarity scores and expected similarity scores, which is calculated as

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}}, \tag{22}$$

where m_x is the mean of vector x and m_y is the mean of vector y. The correlations of $+1$ or -1 suggest an exact linear relationship. The measured correlation was $r = 0.97$ and the p-value was $p = 4.22 \times 10^{-7}$, which indicates that the average similarity scores have a strong positive correlation with the

expected similarity score, thus statistically sufficient. Demo samples can be found online.[1]

V. CONCLUSION

We proposed a CLVC model that is based on the combination of the VAE and StarGAN for controlling speaker individuality. The objective and subjective results indicate that our proposed model, which is trained solely on acoustic features, can effectively control speaker individuality in a cross-lingual setting via the speaker embedding. In terms of over-smoothing, the objective results indicate that our adversarial training scheme can effectively enhance the fine-structure in the converted mel-spectrogram. The results from the subjective test indicate that the improvement in SJ-TE conversion is statistically significant. With the additional language embedding, the language factor can be disentangled from the speaker embedding, avoiding the undesirable effect on linguistic information when converting voice. The objective results also indicated that the F0 injection method can improve the F0 modeling in a CLVC model, which suggests the potential of using modern neural vocoders in a VC model to enhance the quality of converted speech. Moreover, the high correlation between the average similarity score of fictitious voice and the expected similarity score is evidence for a strong linear relation between speaker embedding and perceptual speaker similarity. This finding can be justified for the controllability of speaker individuality in our study.

Our main contribution in this work was to provide an effective model for controlling speaker individuality and several enhancements for CLVC. The results from our study can be directly applied in various applications such as customizing audiobook and avatar voices, dubbing, teleconferencing, singing voice modification, voice restoration after surgery, and cloning of voices of historical persons. In the future, methods for further improving the controllability of speaker individuality will be our next focus.

REFERENCES

[1] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Commun.*, vol. 88, pp. 65–82, Apr. 2017.
[2] T. Nakashika, T. Takiguchi, and Y. Minami, "Non-parallel training in voice conversion using an adaptive restricted Boltzmann machine," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2032–2045, Nov. 2016.
[3] Z. Wu, E. S. Chng, and H. Li, "Conditional restricted Boltzmann machine for voice conversion," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process.*, Jul. 2013, pp. 104–108.
[4] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "AUTOVC: Zero-shot voice style transfer with only autoencoder loss," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 5210–5219.
[5] Y. Sekii, R. Orihara, K. Kojima, Y. Sei, Y. Tahara, and A. Ohsuga, "Fast many-to-one voice conversion using autoencoders," in *Proc. 9th Int. Conf. Agents Artif. Intell.*, 2017, pp. 164–174.
[6] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," in *Proc. Interspeech*, Aug. 2017, pp. 1273–1277.
[7] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Non-parallel voice conversion with cyclic variational autoencoder," in *Proc. Interspeech*, Sep. 2019, pp. 674–678.
[8] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 9, pp. 1432–1443, Sep. 2019.
[9] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2014, pp. 2672–2680.
[10] T. V. Ho and M. Akagi, "Non-parallel voice conversion with controllable speaker individuality using variational autoencoder," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2019, pp. 106–111.
[11] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.
[12] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 266–273.
[13] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *Proc. Interspeech*, Aug. 2017, pp. 3364–3368.
[14] B. Sisman, M. Zhang, M. Dong, and H. Li, "On the study of generative adversarial networks for cross-lingual voice conversion," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2019, pp. 144–151.
[15] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel waveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6199–6203.
[16] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-parallel Sequence-to-Sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, no. 1, pp. 540–552, 2020.
[17] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and D-vectors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5274–5278.
[18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014.
[19] M. Rolinek, D. Zietlow, and G. Martius, "Variational autoencoders pursue PCA directions (by accident)," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12398–12407.
[20] 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. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Dec. 2016.
[21] C. Veaux, J. Yamagishi, and K. Macdonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," Centre Speech Technol. Res., Univ. Edinburgh, Edinburgh, U.K., Tech. Rep., 2017, doi: 10.7488/ds/1994.
[22] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3617–3621.
[23] A. Oord *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 3918–3926.
[24] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, vol. 70, pp. 214–223, 2017.
[25] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: Free Japanese multi-speaker voice corpus," in *Proc. Inf. Process. Soc. Jpn. Res. Rep. (SLP)*, vol. 4, 2019, pp. 1–4.
[26] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. E99.D, no. 7, pp. 1877–1884, 2016.
[27] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *Proc. 9th ISCA Speech Synth. Workshop*, 2016, p. 125.
[28] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016.

[1] https://github.com/tuanvu92/VAE-StarGAN

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.
[30] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 755–767, Apr. 2016.
[31] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 2100–2104.
[32] M. Swain and D. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, pp. 11–32, Nov. 2004.
[33] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through phoneme-based linear mapping functions with straight for mandarin," in *Proc. 4th Int. Conf. Fuzzy Syst. Knowl. Discovery (FSKD)*, vol. 4, 2007, pp. 410–414.

TUAN VU HO received the B.E. degree from Vietnam National University, Ho Chi Minh, in 2015, and the M.E. degree from the Japan Advanced Institute of Science and Technology (JAIST), in 2018, where he is currently pursuing the Ph.D. degree with the Akagi Laboratory. His research interests include voice conversion, speech enhancement, speech-signal processing, and deep learning. He is a member of the Acoustical Society of Japan (ASJ).

MASATO AKAGI (Member, IEEE) received the B.E. degree from the Nagoya Institute of Technology, in 1979, and the M.E. and Ph.D. degrees from the Tokyo Institute of Technology, in 1981 and 1984, respectively.

In 1984, he joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation (NTT). From 1986 to 1990, he worked with the ATR Auditory and Visual Perception Research Laboratories. Since 1992, he has been with the Faculty of the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), where he is currently a Full Professor. His research interests include speech perception, the modeling of speech-perception mechanisms in humans, and the signal processing of speech.