INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Investigation of F0 conditioning and Fully Convolutional Networks in Variational Autoencoder based Voice Conversion

Wen-Chin Huang[12], *Yi-Chiao Wu*[2], *Chen-Chou Lo*[1], *Patrick Lumban Tobing*[2], *Tomoki Hayashi*[2], *Kazuhiro Kobayashi*[2], *Tomoki Toda*[2], *Yu Tsao*[1], *Hsin-Min Wang*[1]

[1] Academia Sinica, Taiwan
[2] Nagoya University, Japan

wen.chinhuang@g.sp.m.is.nagoya-u.ac.jp

Abstract

In this work, we investigate the effectiveness of two techniques for improving variational autoencoder (VAE) based voice conversion (VC). First, we reconsider the relationship between vocoder features extracted using the high quality vocoders adopted in conventional VC systems, and hypothesize that the spectral features are in fact F0 dependent. Such hypothesis implies that during the conversion phase, the latent codes and the converted features in VAE based VC are in fact source F0 dependent. To this end, we propose to utilize the F0 as an additional input of the decoder. The model can learn to disentangle the latent code from the F0 and thus generates converted F0 dependent converted features. Second, to better capture temporal dependencies of the spectral features and the F0 pattern, we replace the frame wise conversion structure in the original VAE based VC framework with a fully convolutional network structure. Our experiments demonstrate that the degree of disentanglement as well as the naturalness of the converted speech are indeed improved.

Index Terms: voice conversion, variational autoencoder, representation disentanglement

1. Introduction

Voice conversion (VC) aims to convert the speech from a source to that of a target without changing the linguistic content. Numerous approaches have been proposed, such as Gaussian mixture model (GMM)-based methods [1, 2], deep neural network (DNN)-based methods [3, 4], and exemplar-based methods [5, 6, 7]. Most of them require parallel training data, i.e., the source and target speakers utter the same transcripts for training. Since such data is hard to collect, non-parallel training has long remained one of the ultimate goals in VC.

Recently, VAEs [8] have been successfully applied to VC [9], which we will refer to as VAE-VC. Specifically, the spectral conversion function is composed of an encoder-decoder pair. The encoder first encodes the input spectral feature into a latent code. Then, the decoder mixes the latent code and the target speaker code to generate the output. The encoder-decoder network and the speaker codes are trained by back-propagation of the reconstruction error, along with a Kullback-Leibler (KL)-divergence loss that regularizes the distribution of the latent variable, thus there is no need for parallel training data. The success of this framework implies that the encoder learns to eliminate the speaker dependent information from the input, making the latent code speaker independent.

Following conventional VC systems [10], the VAE-VC framework first utilizes high quality vocoders such as WORLD [11] and STRAIGHT [12] to extract different kinds of acoustic features, e.g., spectral feature and fundamental frequency (F0). As depicted in Fig. 1, these features are then converted separately, and a waveform synthesizer finally generates the converted waveform using the converted features. The validity of converting acoustic features in different feature streams comes from the assumption that these features are independent from each other. However, during feature extraction, both WORLD and STRAIGHT extract the F0 first, then use the extracted F0 to obtain the spectral feature. Thus, we hypothesize that the spectral features are in fact F0 dependent.

Based on this hypothesis, we may imply that during the conversion phase in VAE-VC, since the encoder was never trained to eliminate F0 information, the latent code extracted from the source spectral feature still contains information of the source F0, thus the converted feature is source F0 dependent. It can be assumed that the conversion performance can suffer from this flaw. In other words, if the converted spectral feature can be made converted F0 dependent, the performance will improve. This is analogous to a previous work that models the cross stream dependency in Hidden Markov Model based speech synthesis [13, 14]

How do we obtain converted spectral features that are converted F0 dependent? In VAE-VC, by conditioning the decoder with the speaker code, the encoder can learn to eliminate speaker dependent information from the input, thus make the latent code speaker independent. We may assume that, similarly, by conditioning the decoder with F0, we may disentangle the latent code from F0, as illustrated in Fig. 2. As a result, during the conversion phase, given the converted F0, we may obtain the desired converted F0 dependent converted spectral feature.

However, applying the above mentioned concept to VAE-VC may be somehow problematic. The original VAE-VC performs conversion in a frame wise manner, i.e., no temporal relationship is considered. Under such structure, if the model is given an F0 value per frame, it is concerned whether the encoder can actually learn to eliminate F0 information effectively. Therefore, we assume that by designing the network to be able to acquire an input sequence, the model can benefit from capturing the F0 contour, and thus better disentangle the latent code.

In this work, we investigate two techniques to improve the general VAE-VC framework. Specifically, we condition the decoder with F0, and adopt the fully convolutional network (FCN) [15] to consider temporal dependencies of the inputs. Our contributions are:

- We reconsidered the relationship between different vocoder features, and hypothesized that the use of F0 as an additional condition variable of the model can eliminate the source F0 information in the encoded latent codes, and as a result obtain the converted F0 dependent

http://dx.doi.org/10.21437/Interspeech.2019-1774

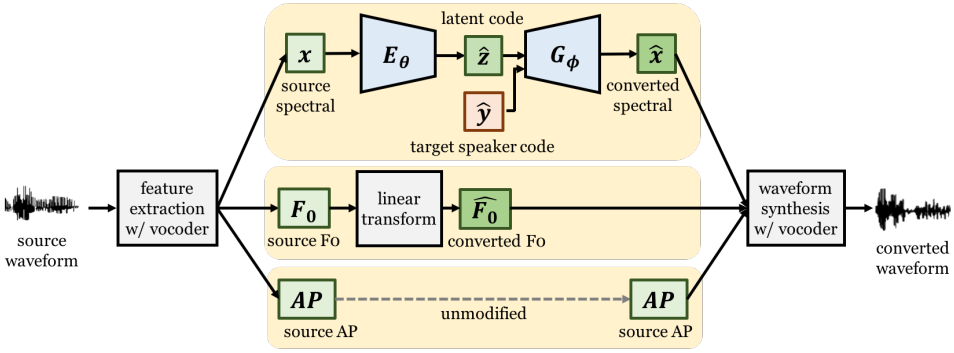


Figure 1: *Illustration of the VAE-VC framework. Following traditional VC systems, a vocoder first parameterizes the waveform into acoustic features, which are then converted in different streams, and finally the converted features are used to synthesize the converted waveform by a vocoder.*

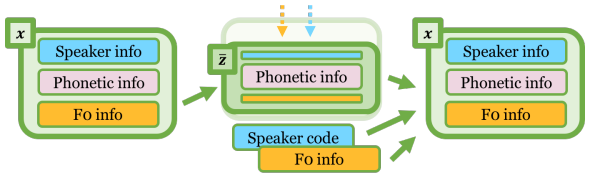


Figure 2: *Disentangling the latent code with condition variables. By providing the speaker code and F0 explicitly, the encoder learns to discard as much speaker and F0 information as possible, thereby generating compact latent representations.*

converted spectral feature. We verified this hypothesis by showing that the latent code obtained in this way is indeed more F0 independent through objective measures, and evaluated the conversion performance.

- We adopted the FCN structure so that the model can take an input sequence in order to consider the temporal relationship of the spectral features and F0 pattern. Note that the FCN structure was first combined with VAE-VC in [16], but the impact of FCN was not solely examined. We provide detailed experiments to examine the effectiveness of this structure.

2. Related work on VAE-VC

2.1. VAE-VC

Figure 1 illustrates a VAE-VC system [9]. The core of VAE-VC is an encoder-decoder network, which models the WORLD spectra (SP). During training, given an input spectral frame $\boldsymbol{x}$, the encoder E_θ with parameter set θ encodes $\boldsymbol{x}$ into a latent code: $\boldsymbol{z} = E_\theta(\boldsymbol{x})$. The speaker code $\boldsymbol{y}$ of the input frame, along with $\boldsymbol{z}$, are passed to the decoder G_ϕ with parameter set ϕ to reconstruct the input. This reconstruction process can be written as:

$$\bar{\boldsymbol{x}} = G_\phi(\boldsymbol{z}, \boldsymbol{y}) = G_\phi(E_\theta(\boldsymbol{x}), \boldsymbol{y}). \tag{1}$$

The model parameters can be obtained by maximizing the variational lower bound:

$$\mathcal{L}_{vae}(\theta, \phi; \boldsymbol{x}, \boldsymbol{y}) = \mathcal{L}_{recon}(\boldsymbol{x}, \boldsymbol{y}) + \mathcal{L}_{lat}(\boldsymbol{x}), \tag{2}$$

$$\mathcal{L}_{recon}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{\boldsymbol{z}\sim q_\theta(\boldsymbol{z}|\boldsymbol{x})}\big[\log p_\phi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y})\big], \tag{3}$$

$$\mathcal{L}_{lat}(\boldsymbol{x}) = -D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})), \tag{4}$$

where $q_\theta(\boldsymbol{z}|\boldsymbol{x})$ is the approximate posterior, $p_\phi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y})$ is the data likelihood, and $p(\boldsymbol{z})$ is the prior distribution of the latent space. $\mathcal{L}_{recon}$ is simply a reconstruction term as in any vanilla auto encoder, whereas $\mathcal{L}_{lat}$ regularizes the encoder to align the approximate posterior with the prior distribution.

In the conversion phase, one could use (1) to formulate the conversion function f with the target speaker $\hat{\boldsymbol{y}}$:

$$\hat{\boldsymbol{x}} = f(\boldsymbol{x}, \hat{\boldsymbol{y}}) = G_\phi(\boldsymbol{z}, \hat{\boldsymbol{y}}) = G_\phi(E_\theta(\boldsymbol{x}), \hat{\boldsymbol{y}}). \tag{5}$$

There is a line of work extending the VAE-VC framework. [17] was the first VC framework to incorporate generative adversarial network (GAN) to improve spectral modeling. [18] utilized external modules such as automatic speech recognition and speaker verification systems to obtain phonetic posteriorgrams and d-vectors to improve the performance. [16] borrowed the idea of auxiliary classifiers from conditional image generation to force the decoder to preserve more speaker characteristics, and further used an FCN structure to take sequential input features.

In the following subsection, we will introduce the baseline system we use in this paper.

2.2. CDVAE-VC

[19] proposed a cross-domain VAE framework, by extending the conventional VAE framework to jointly consider two kinds of spectral features, SPs and Mel-Cepstral Coefficients (MCCs), extracted from the same observed speech frame, to utilize their different properties. This framework, which we will refer to as CDVAE-VC, is a collection of encoder-decoder pairs, one for each kind of spectral feature. Considering the SPs and MCCs as two kinds of spectral features (denoted as $\boldsymbol{x}_{SP}$ and $\boldsymbol{x}_{MCC}$), we define the following losses:

$$\boldsymbol{z}_{SP} = E_{SP}(\boldsymbol{x}_{SP}), \boldsymbol{z}_{MCC} = E_{MCC}(\boldsymbol{x}_{MCC}), \tag{6}$$

$$\bar{\boldsymbol{x}}_{S-S} = G_{SP}(\boldsymbol{z}_{SP}, \boldsymbol{y}), \bar{\boldsymbol{x}}_{M-M} = G_{MCC}(\boldsymbol{z}_{MCC}, \boldsymbol{y}), \tag{7}$$

$$\bar{\boldsymbol{x}}_{S-M} = G_{MCC}(\boldsymbol{z}_{SP}, \boldsymbol{y}), \bar{\boldsymbol{x}}_{M-S} = G_{SP}(\boldsymbol{z}_{MCC}, \boldsymbol{y}), \tag{8}$$

$$\mathcal{L}_{in} = \mathcal{L}_{recon}(\bar{\boldsymbol{x}}_{S-S}, \boldsymbol{y}) + \mathcal{L}_{recon}(\bar{\boldsymbol{x}}_{M-M}, \boldsymbol{y}), \tag{9}$$

$$\mathcal{L}_{KLD} = \mathcal{L}_{lat}(\boldsymbol{x}_{SP}) + \mathcal{L}_{lat}(\boldsymbol{x}_{MCC}), \tag{10}$$

$$\mathcal{L}_{cross} = \mathcal{L}_{recon}(\bar{\boldsymbol{x}}_{S-M}, \boldsymbol{y}) + \mathcal{L}_{recon}(\bar{\boldsymbol{x}}_{M-S}, \boldsymbol{y}), \tag{11}$$

$$\mathcal{L}_{sim} = \|\boldsymbol{z}_{SP} - \boldsymbol{z}_{MCC}\|_1, \tag{12}$$

where E_{SP} and E_{MCC} are the encoders for SP and MCC, respectively, while G_{SP} and G_{MCC} are decoders.

In short, two extra reconstruction streams were introduced. By optimizing the cross-domain reconstruction loss, $\boldsymbol{z}_{SP}$ is enforced to contain enough information to reconstruct $\boldsymbol{x}_{MCC}$, and

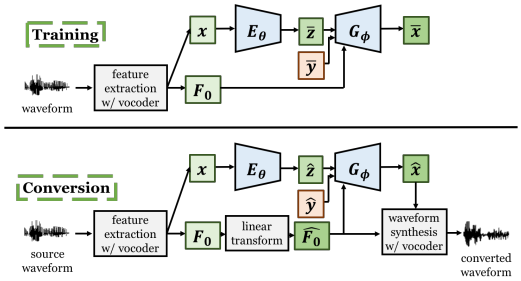


Figure 3: *The proposed framework with F0 conditioning.*

vice versa. As a result, the behavior of the encoders from both feature domains are constrained to be the same, i.e., they are expected to extract similar latent information from different types of input spectral features. To explicitly reinforce this constraint, a latent similarity loss was also included.

The final objective is as follows:

$$\mathcal{L}_{cdvae} = \mathcal{L}_{in} + \mathcal{L}_{KLD} + \mathcal{L}_{cross} + \mathcal{L}_{sim}. \quad (13)$$

The model parameters can be learned by minimizing (13). In the conversion phase, there are four conversion paths (i.e., two within-domain and two cross-domain paths). As reported in [19], the MCC-MCC path gave the best performance in terms of subjective measure, which matched the common assumption that MCCs are related to human perception.

2.3. Problem definition

Here we once again point out the flaws of the general VAE-VC framework. First, since the model is built on a frame wise basis, the network is limited in modeling the temporal dependencies of speech. Second, it is possible that the spectral features extracted using a vocoder is actually F0 dependent. Thus, F0 information in the source spectral feature might remain in the encoded latent code as well as the converted spectral feature, thereby damaging the conversion performance.

3. Investigated Methods

In this section, we examine two mechanisms to overcome the disadvantages in VAE-VC mentioned in Section 2.3.

3.1. Modeling time dependencies with the FCN structure

When it comes to sequential models, the recurrent neural network (RNN) is a commonly chosen network structure. Nonetheless, we follow [16] and adopt the FCN structure. There are several reasons why we choose FCNs over RNNs. First, the nature of RNNs introduces high computational costs, and convolutional layers make parallel computation feasible. Second, RNNs have an infinitely large receptive field in theory, and we think this is unnecessary in our task. In contrast, by adjusting the depth and kernel sizes, convolutional neural networks can be flexibly designed to have a finite, reasonably large receptive field. Note that here the output of our model is still of the same length as the input. Although sequence to sequence based models, which can generate output sequences of variable length, have been successfully applied to VC [20, 21, 22, 23, 24], we will show that only considering temporal dependencies can bring significant improvements to VAE-VC.

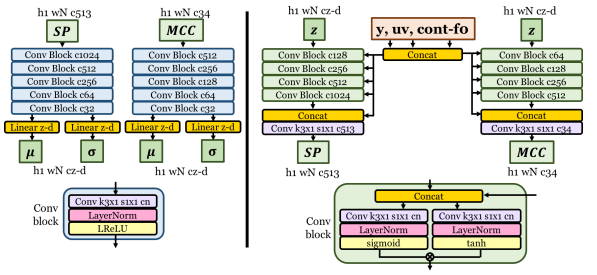


Figure 4: *Model architecture. The input is of length N. h, w, c means height, width and channels. z-d means latent code dimension. LReLU, LayerNorm means leaky rectified activation function and layer normalization layer. k, s, c means kernel size, stride and output channels.*

3.2. Conditioning on F0

We propose to use F0 as an additional condition variable in order to eliminate the F0 information in the latent code, as shown in Fig. 3. Specifically, during training, given the F0 contour of the input F_0, we modify (1) as:

$$\bar{\boldsymbol{x}} = G_\phi(\boldsymbol{z}, \boldsymbol{y}, F_0) = G_\phi(E_\theta(\boldsymbol{x}), \boldsymbol{y}, F_0). \quad (14)$$

During conversion, given the converted F0 contour $\hat{F}_0$, we modify (14) to obtain:

$$\hat{\boldsymbol{x}} = f(\boldsymbol{x}, \hat{\boldsymbol{y}}, \hat{F}_0) = G_\phi(\boldsymbol{z}, \hat{\boldsymbol{y}}, \hat{F}_0) = G_\phi(E_\theta(\boldsymbol{x}), \hat{\boldsymbol{y}}, \hat{F}_0). \quad (15)$$

In our preliminary experiments, we tested several combinations of prosodic features, including *F0*, *continuous (interpolated, cont) F0 + uv symbol* and *cont F0 + uv + band APs*, and found that *cont F0 + uv* had the best performance.

4. Experimental Evaluation

4.1. Experimental settings

We evaluated our proposed methods on the Voice Conversion Challenge 2018 dataset [25], which included recordings of professional US English speakers with a sampling rate of 22050 Hz. The dataset consisted of 81/35 utterances per speaker for training/testing sets, respectively. We used the first 70 utterances of the training set of all speakers for training, the remaining 11 for validation, and the 35 in the testing set of speakers SF1, SF2, SM1, SM2, TF1, TF2, TM1, TM2 to form 16 conversion pairs for evaluation. The WORLD vocoder [11] was adopted to extract acoustic features including 513-dimensional SPs, 513-dimensional APs and F0. The SPs were normalized to unit-sum, and the normalizing factor was taken out and thus not modified. 35-dimensional MCCs were further extracted from the SPs. In the conversion phase, the energy and APs were kept unmodified, and the F0 was converted using a linear mean-variance transformation in the log domain.

The baseline system was the CDVAE-VC [19] system. On top of CDVAE, we first replace the original frame wise structure with FCNs to take sequential inputs. We will refer to this model as FCN-CDVAE. The architecture, as illustrated in Fig. 4, was very similar to [16], using gated linear units activation function and skip connections in the decoders to better propagate the conditional information. Following [26, 27], we randomly sampled 128 frames with overlap during training. For both frame wise and sequential models, the batch size, latent dimension and speaker code dimension were all 16, and the models were

Table 1: *Mean Mel-cepstral distortion [dB] of all non-silent frames from the baseline and proposed models.*

System	F-F	F-M	M-F	M-M	Avg.
CDVAE	6.67	6.31	6.71	5.97	6.42
FCN-CDVAE	6.57	6.27	6.97	5.76	6.39
F0-FCN-CDVAE	6.56	6.31	6.86	5.79	6.38

Table 2: *MOS for naturalness with 95% confidence intervals.*

System	Avg.
CDVAE	2.45 ± 0.13
FCN-CDVAE	2.89 ± 0.15
F0-FCN-CDVAE	2.84 ± 0.17
Target	4.96 ± 0.04

trained using the Adam optimizer [28] with learning rate, β_1 and β_2 set to 0.0001, 0.5, and 0.999. Note that the speaker codes were randomly initialized and optimized, as in [19].

We applied the F0 conditioning mechanism to the FCN-CDVAE model, referred to as F0-FCN-CDVAE. We concatenated these features with the speaker code in the feature axis.

4.2. Effectiveness of FCN

We first examined the effectiveness of FCN. Table 1 shows the mean Mel-cepstral distortion (MCD) values, and Table 2 shows the mean opinion scores (MOS) obtained from a listening test where 10 participants were asked to evaluate the naturalness of the speech on a five-point scale, including the natural target speech. We may conclude from the results that the FCN structure indeed improved the overall performance in terms of both MCD and MOS, except for the M-F conversion pairs, where we will leave the investigation for future work.

4.3. Effect of F0 conditioning

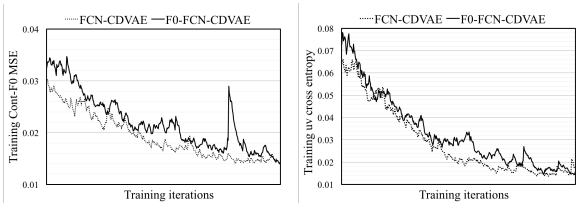
We then examined the F0 conditioning mechanism. We conducted several experiments to check if the latent codes are indeed more disentangled from F0. First, consider two sentences with the same content uttered by the source and the target speakers. An ideal encoder should extract identical latent codes from these two sentences since the phonetic contents are the same, though with different styles such as F0. We measured the distance between latent code pairs in terms of root mean squared error (RMSE) and cosine similarity, as in Table 3. The results showed that the latent code pairs extracted using F0-FCN-CDVAE were closer in terms of RMSE and cosine similarity.

The amount of F0 information that resides in the latent code also reflects the degree of disentanglement. Following [29], we trained an F0 prediction network and reported the training loss. Specifically, a network with the same architecture as the encoder in Fig. 4 was trained to take a sequence of latent codes as input and predict the corresponding *cont-F0* + *uv*. We assumed that less F0 information left in the latent codes results in worse training performance. Fig. 5a and Fig. 5b show the training *cont-F0* mean squared error (MSE) and *uv* cross entropy. As expected, the training losses with latent codes extracted using F0-FCN-CDVAE were higher, suggesting less F0 information present in the latent codes.

We finally examined the performance of VC. As reported in Tables 1 and 2, F0-FCN-CDVAE had an similar performance compared with FCN-CDVAE in terms of MCD and MOS.

Table 3: *RMSE distance and cosine similarity of latent codes extracted from parallel sentences of source-target pairs over non-silent frames.*

System	F-F	F-M	M-F	M-M	Avg.
[RMSE]					
FCN-CDVAE	.335	.337	.353	.304	.333
F0-FCN-CDVAE	.284	.286	.302	.260	.283
[Cosine Similarity]					
FCN-CDVAE	.530	.502	.475	.588	.524
F0-FCN-CDVAE	.579	.547	.519	.616	.565



(a) (b)

Figure 5: *F0 prediction network training loss of (a) Cont-F0 MSE (b) uv cross entropy.*

5. Discussions and Conclusions

In this work, we investigated two approaches to improve VAE-VC: an FCN structure to capture the temporal relationship of speech, and the F0 conditioning mechanism to eliminate residual F0 information in the latent code that might potentially harm the conversion performance. The experimental evaluations showed that the impact of FCNs on the objective measures and subjective speech naturalness assessment was positive. On the other hand, F0 conditioning showed promising results in increasing the degree of disentanglement of latent codes, and achieved high speech quality equivalent to FCN-CDVAE. Speech samples are available at [30].

We attribute the insignificant improvement brought by the F0 conditioning scheme to a mismatch between training and conversion. The converted F0 obtained through such a simple F0 conversion process adopted in this and many past works is far from natural. As a result, in the conversion phase, the input combination which consisted of latent codes extracted from normal MCCs and the unnatural converted F0 might have not be seen by the model during training, thereby causing a degradation in quality.

Despite the above mentioned mismatch, we would like to highlight that, although the motivation of applying F0 conditioning to VAE-VC was based on the assumption that the vocoder spectral features are F0 dependent, the design of vocoders was to *separate* these two features as much as possible. The amount of F0 information that resides in the spectral features might already be small enough for our proposed mechanism to eliminate. In the future, we plan to apply this general idea to rawer input features that are richer in F0, e.g., magnitude spectrograms.

6. Acknowledgements

This work was partly supported by JSPS KAKENHI Grant Numbers JP17H06101 and 17H01763, as well as MOST-Taiwan Grants 105-2221-E001-012-MY3 and 107-2221-E-001-008-MY3.

7. References

[1] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar 1998.

[2] T. Toda, A. W. Black, and 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, pp. 2222–2235, Nov 2007.

[3] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 954–964, July 2010.

[4] L. H. Chen, Z. H. Ling, L. J. Liu, and L. R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1859–1872, Dec 2014.

[5] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Proc. SLT*, 2012, pp. 313–317.

[6] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1506–1521, Oct 2014.

[7] Y.-C. Wu, H.-T. Hwang, C.-C. Hsu, Y. Tsao, and H.-M. Wang, "Locally linear embedding for exemplar-based spectral conversion," in *Proc. Interspeech*, 2016, pp. 1652–1656.

[8] D. P. Kingma and M. Welling, "Auto-encoding variational bayes." *CoRR*, vol. abs/1312.6114, 2013.

[9] 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. APISPA ASC*, 2016, pp. 1–6.

[10] K. Kobayashi, S. Takamichi, S. Nakamura, and T. Toda, "The nunaist voice conversion system for the voice conversion challenge 2016," in *Interspeech 2016*, 2016, pp. 1667–1671. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-970

[11] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Transactions on Information and Systems*, vol. 99, pp. 1877–1884, 2016.

[12] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187 – 207, 1999.

[13] Z. Ling, W. Zhang, and R. Wang, "Cross-stream dependency modeling for hmm-based speech synthesis," in *2008 6th International Symposium on Chinese Spoken Language Processing*, Dec 2008, pp. 1–4.

[14] X. Wang, Z. Ling, and L. Dai, "Cross-stream dependency modeling using continuous f0 model for hmm-based speech synthesis," in *2012 8th International Symposium on Chinese Spoken Language Processing*, Dec 2012, pp. 84–87.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3431–3440.

[16] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Acvae-vc: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder," *arXiv e-prints*, p. arXiv:1808.05092, Aug 2018.

[17] 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*, 2017, pp. 3364–3368.

[18] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5274–5278.

[19] W.-C. Huang, H.-T. Hwang, Y.-H. Peng, Y. Tsao, and H.-M. Wang, "Voice conversion based on cross-domain features using variational auto encoders," in *Proc. ISCSLP*, 2018.

[20] H. Miyoshi, Y. Saito, S. Takamichi, and H. Saruwatari, "Voice conversion using sequence-to-sequence learning of context posterior probabilities," in *Proc. Interspeech 2017*, 2017, pp. 1268–1272. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-247

[21] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," in *Proc. Interspeech 2017*, 2017, pp. 1283–1287. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-970

[22] J.-X. Zhang, Z.-H. Ling, L.-R. Dai, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *arXiv e-prints*, p. arXiv:1810.06865, Oct 2018.

[23] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," *CoRR*, vol. abs/1811.04076, 2018.

[24] H. Kameoka, K. Tanaka, T. Kaneko, and N. Hojo, "Convs2s-vc: Fully convolutional sequence-to-sequence voice conversion," *CoRR*, vol. abs/1811.01609, 2018.

[25] J. Lorenzo-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, pp. 195–202.

[26] T. Kaneko and H. Kameoka, "Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks," *ArXiv e-prints*, Nov. 2017.

[27] J. chieh Chou, C. chieh Yeh, H. yi Lee, and L. shan Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," in *Proc. Interspeech 2018*, 2018, pp. 501–505. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1830

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[29] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *CoRR*, vol. abs/1901.08810, 2019.

[30] https://unilight.github.io/Publication-Demos/publications/f0-fcn-cdvae/, accessed: 2019-07-01.