EasyChair Preprint

№ 103

NU Voice Conversion System for the Voice Conversion Challenge 2018

Patrick Lumban Tobing, Yichiao Wu, Tomoki Hayashi,
Kazuhiro Kobayashi and Tomoki Toda

April 28, 2018

NU Voice Conversion System for the Voice Conversion Challenge 2018

Patrick Lumban Tobing[1]*, Yi-Chiao Wu*[1]*, Tomoki Hayashi*[1]*, Kazuhiro Kobayashi*[2]*, Tomoki Toda*[2]

[1]Graduate School of Information Science, Nagoya University, Japan
[2]Information Technology Center, Nagoya University, Japan

{patrick.lumbantobing, yichiao.wu, hayashi.tomoki}@g.sp.m.is.nagoya-u.ac.jp
kobayashi.kazuhiro@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

Abstract

This paper presents the NU (Nagoya University) voice conversion (VC) system for the HUB task of the Voice Conversion Challenge 2018 (VCC 2018). The design of the NU VC system can basically be separated into two modules consisting of a speech parameter conversion module and a waveform-processing module. In the speech parameter conversion module, a deep learning framework is deployed to estimate the spectral parameters of a target speaker given those of a source speaker. Specifically, a deep neural network (DNN) and a deep mixture density network (DMDN) are used as the deep model structure. In the waveform-processing module, given the estimated spectral parameters and linearly transformed F0 parameters, the converted waveform is generated using a WaveNet-based vocoder system. To use the WaveNet-based vocoder, there are several generation flows based on an analysis-synthesis framework to obtain the speech parameter set, on the basis of which a system selection process is performed to select the best one in an utterance-wise manner. The results of VCC 2018 ranked the NU VC system in second place with an overall mean opinion score (MOS) of 3.44 for speech quality and 85% accuracy for speaker similarity.

1. Introduction

Every human being has their own speech characteristics. The capability of handling the speaker characteristics within a speech signal has great potential to be employed in real-world applications. Indeed, this so-called voice conversion (VC) framework has been used in several works, such as, singing voice conversion [1, 2], body-conducted speech conversion [3], speech signal recovery [4, 5], and speech modification [6]. The growing interest in VC development motivated many researchers around the world to conceive the 1st Voice Conversion Challenge in 2016 [7]. Following this, we participated in the 2nd Challenge, i.e., the Voice Conversion Challenge (VCC) 2018 [8]. Our participating system, the NU (Nagoya University) VC system, is elaborated in this paper.

In the development of a VC system, three aspects need to be considered: the conversion of spectral parameters, the conversion of prosodic parameters, and waveform generation. In the spectral parameter conversion, many techniques based on statistical methods have been proposed, such as, codebook-based conversion [9], Gaussian mixture model (GMM)-based mapping [10], and a neural-network based system [11, 12]. On the other hand, in the handling of prosodic parameters, such as fundamental frequency (F0), several methods have been commonly used including a simple mean/variance linear transformation, a contour-based transformation [13], GMM-based mapping [14], and neural network [15]. For waveform generation, approaches include the source-filter vocoder system [16], the latest direct waveform modification technique [2], and the use of state-of-the-art WaveNet modeling [17, 18, 19].

In this paper, we describe the NU VC system, which uses a neural network architecture for spectral modeling as well as a WaveNet-based vocoder for waveform modeling and generation. The NU VC system adopts a neural network design for spectral parameter conversion, where a structure combining a deep neural network (DNN) and a deep mixture density network (DMDN) [20] is used to form a cascaded DMDN (CascDMDN). In a conventional DNN or DMDN, given a sequence of source spectral parameters, the target sequence is estimated using a single Gaussian distribution in a DNN or using a mixture of Gaussian distributions in a DMDN. In CascDMDN, a sequence of estimated source spectral parameters is first inferred within its first set of hidden layers, which is then fed into the second set to estimate the target sequence. For the conversion of prosodic parameters, the NU VC system uses a linear transformation of framewise F0 values of the source speaker into those of the target on the basis of their mean and variance statistics.

In the waveform-processing module, the NU VC system deploys the state-of-the-art WaveNet-based vocoder [17, 18, 19] framework to directly model the waveform. In WaveNet [21], each waveform sample is conditioned using previous samples and possible auxiliary features within a stack of dilated convolutional layers. The structure of the dilated convolutions makes it possible to exponentially increase the receptive field of waveform samples efficiently. In addition, in the NU VC system, the auxiliary features include the voiced/unvoiced (U/V) decision, continuous F0 values, mel-cepstrum parameters, and aperiodicity features. To obtain the set of refined speech parameters, the NU VC system carries out direct waveform modification [2] in several analysis-synthesis flows. Then, a model selection procedure is performed to select the best waveform generation flow in an utterance-wise manner. In the evaluations carried out at the VCC 2018, the NU VC system achieved second place with an average mean opinion score (MOS) of 3.44 for speech quality and 85% accuracy for speaker similarity.

The rest of the paper is organized as follows. Spectral parameter conversion models are elaborated in Section 2. The waveform-processing module is described in Section 3. Experimental results are presented in Section 4. Finally, the conclusion is given in Section 5.

2. Spectral parameter conversion models

In this section, the deep learning structures used to perform spectral parameter conversion are elaborated. Their graphical model representations are illustrated in Fig. 1. Moreover, the

overall process described in this section is illustrated in the upper diagram of Fig. 2.

2.1. Conversion model with deep neural network (DNN)

Let $\boldsymbol{x}_t = [x_t(1), x_t(2), \ldots, x_t(D)]^\top$ and $\boldsymbol{y}_t = [y_t(1), y_t(2), \ldots, y_t(D)]^\top$ be the D-dimensional spectral feature vector of the source speaker and that of the target speaker at frame t, respectively. The $2D$-dimensional joint static-delta feature vector of the source and that of the target are then respectively denoted as $\boldsymbol{X}_t = [\boldsymbol{x}_t, \Delta\boldsymbol{x}_t]^\top$ and $\boldsymbol{Y}_t = [\boldsymbol{y}_t, \Delta\boldsymbol{y}_t]^\top$ at frame t, where the delta feature vectors are denoted as $\Delta\boldsymbol{x}_t$ and $\Delta\boldsymbol{y}_t$.

In the conventional DNN architecture, given an input source spectral feature vector $\boldsymbol{X}_t$ and the network parameters $\boldsymbol{\lambda}$, a conditional probability distribution function (pdf) of the target spectral feature vector $\boldsymbol{Y}_t$ on the network output layer is defined as follows:

$$P_s(\boldsymbol{Y}_t|\boldsymbol{X}_t, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{Y}_t; f_\lambda(\boldsymbol{X}_t), \boldsymbol{D}), \tag{1}$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. In the above pdf, the network output is denoted as $f_\lambda(\boldsymbol{X}_t)$ and the diagonal covariance matrix of the target spectral feature vector is denoted as $\boldsymbol{D}$, which is inferred from training data. The DNN spectral conversion model is represented by the left graph in Fig. 1.

In the training phase, a set of updated network parameters $\hat{\boldsymbol{\lambda}}$ is estimated by backpropagating the following loss function:

$$\hat{\boldsymbol{\lambda}} = \operatorname*{argmin}_{\boldsymbol{\lambda}} -P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}), \tag{2}$$

where

$$P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) = \prod_{t=1}^{T} P_s(\boldsymbol{Y}_t|\boldsymbol{X}_t, \boldsymbol{\lambda}). \tag{3}$$

The spectral feature vector sequence of the source speaker and that of the target speaker are denoted as $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \ldots, \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, \ldots, \boldsymbol{Y}_T^\top]^\top$, respectively. Note that in the training phase, a dynamic time warping (DTW) procedure is performed by aligning the length of the source spectral feature vector sequence with that of the target one to obtain a pair of time-aligned features.

In the conversion phase, given the source spectral feature vector sequence $\boldsymbol{X}$, the trajectory of the target spectral parameters $\hat{\boldsymbol{y}} = [\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, \ldots, \boldsymbol{y}_T^\top]^\top$ is computed by the maximum likelihood parameter generation (MLPG) [22] procedure as follows:

$$\hat{\boldsymbol{y}} = (\boldsymbol{W}^\top \boldsymbol{U}^{-1} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{U}^{-1} \boldsymbol{M}, \tag{4}$$

where $\boldsymbol{W}$ is a transformation matrix used to expand a static feature vector sequence into its joint static-delta feature vector sequence. The sequence of target mean vectors is denoted as $\boldsymbol{M} = [f_\lambda(\boldsymbol{X}_1)^\top, f_\lambda(\boldsymbol{X}_2)^\top, \ldots, f_\lambda(\boldsymbol{X}_T)^\top]^\top$, whereas the sequence of diagonal covariance matrices is denoted as $\boldsymbol{U} = \boldsymbol{D} \otimes \boldsymbol{I}_{2D \times T}$ with $\otimes$ denoting the Kronecker delta product.

2.2. Conversion model with deep mixture density network (DMDN)

The NU VC system uses a DMDN [20] in the spectral parameter conversion by inferring a mixture of pdfs of the target spectral feature vector. Given an input source feature vector $\boldsymbol{X}_t$ at frame t, the conditional pdf of the target spectral feature vector $\boldsymbol{Y}_t$ is

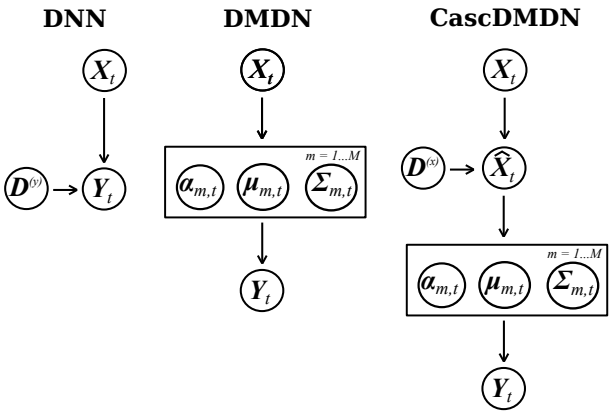

Figure 1: Graphical representations of spectral conversion models using DNN, DMDN, and CascDMDN.

then defined as follows:

$$P_m(\boldsymbol{Y}_t|\boldsymbol{X}_t, \boldsymbol{\lambda}) = \sum_{m=1}^{M} \alpha_{m,t} P(\boldsymbol{Y}_t|\boldsymbol{\mu}_{m,t}, \boldsymbol{\Sigma}_{m,t}), \tag{5}$$

where the time-varying target mean vector and diagonal covariance matrix are respectively denoted as $\boldsymbol{\mu}_{m,t}$ and $\boldsymbol{\Sigma}_{m,t}$ for the mth mixture component. The weight of the mth mixture component is denoted as $\alpha_{m,t}$. The total number of mixture components is M. These time-varying mixture parameters are taken from the network output $f_\lambda(\boldsymbol{X}_t) = [f_\lambda^{(\alpha_1)}(\boldsymbol{X}_t), f_\lambda^{(\mu_1)}(\boldsymbol{X}_t)^\top, f_\lambda^{(\Sigma_1)}(\boldsymbol{X}_t)^\top, \ldots, f_\lambda^{(\alpha_M)}(\boldsymbol{X}_t), f_\lambda^{(\mu_M)}(\boldsymbol{X}_t)^\top, f_\lambda^{(\Sigma_M)}(\boldsymbol{X}_t)^\top]^\top$ as

$$\alpha_{m,t} = \frac{f_\lambda^{(\alpha_m)}(\boldsymbol{X}_t)}{\sum_{n=1}^{M} f_\lambda^{(\alpha_n)}(\boldsymbol{X}_t)} \tag{6}$$

$$\boldsymbol{\mu}_{m,t} = f_\lambda^{(\mu_m)}(\boldsymbol{X}_t) \tag{7}$$

$$\boldsymbol{\Sigma}_{m,t} = \operatorname{diag}[\exp(f_\lambda^{(\Sigma_m)}(\boldsymbol{X}_t))^{\circ 2}], \tag{8}$$

where $\circ$ denotes a Hadamard elementwise product. The DMDN spectral conversion model is represented by the middle graph in Fig. 1.

In the training phase, a set of updated network parameters $\hat{\boldsymbol{\lambda}}$ is estimated by backpropagating the negative log likelihood derived from the conditional pdf given in Eq. (5) in a similar manner to the DNN in Eq. (2). On the other hand, in the conversion phase using the DMDN, given a source spectral feature vector sequence $\boldsymbol{X}$, the trajectory of the target spectral parameters $\hat{\boldsymbol{y}}$ is estimated by also using the MLPG [22] procedure as follows:

$$\hat{\boldsymbol{y}} = (\boldsymbol{W}^\top \overline{\boldsymbol{\Sigma}}^{-1} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \overline{\boldsymbol{\Sigma}}^{-1} \overline{\boldsymbol{\mu}}. \tag{9}$$

The sequence of the target mean vectors and that of the diagonal covariance matrices in the above equation are respectively given by

$$\overline{\boldsymbol{\mu}} = [\boldsymbol{\mu}_{\hat{m}_1,1}^\top, \boldsymbol{\mu}_{\hat{m}_2,2}^\top, \ldots, \boldsymbol{\mu}_{\hat{m}_T,T}^\top]^\top \tag{10}$$

$$\overline{\boldsymbol{\Sigma}} = \operatorname{diag}[\boldsymbol{\Sigma}_{\hat{m}_1,1}, \boldsymbol{\Sigma}_{\hat{m}_2,2}, \ldots, \boldsymbol{\Sigma}_{\hat{m}_T,T}], \tag{11}$$

where the suboptimum mixture component sequence $\hat{\boldsymbol{m}} = \{\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_T\}$ is determined as follows:

$$\hat{\boldsymbol{m}} = \operatorname*{argmax}_{\boldsymbol{m}} \prod_{t=1}^{T} \alpha_{m_1,t}. \tag{12}$$

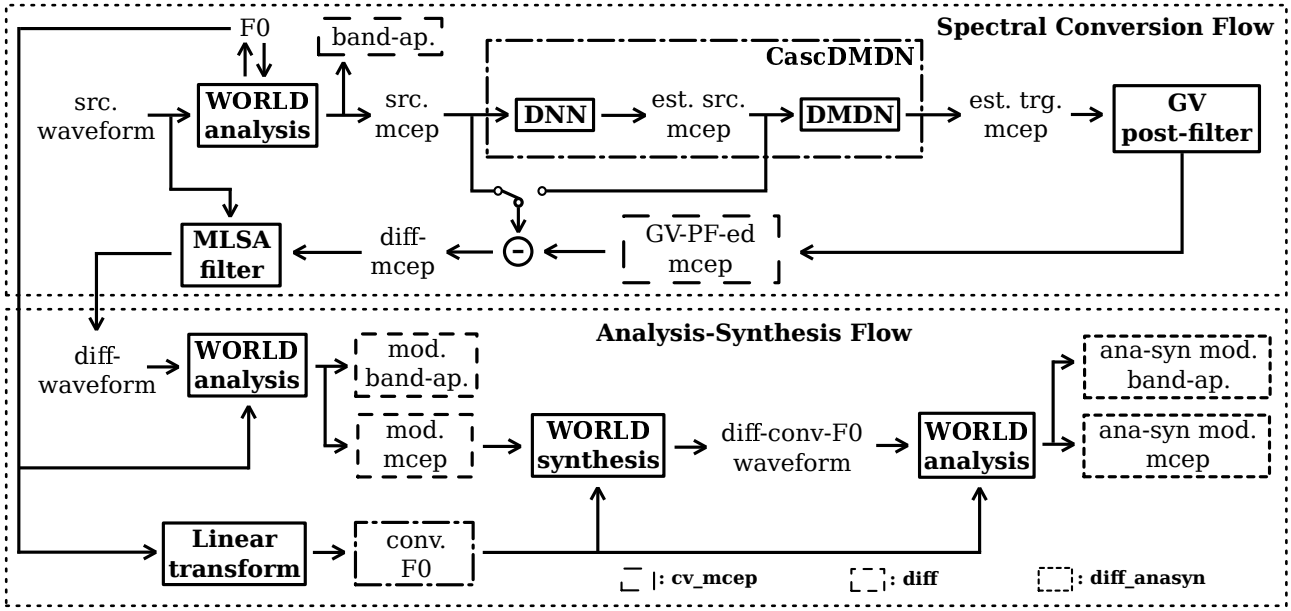

Figure 2: Diagram of the spectral conversion flow (top) and the analysis-synthesis flow (bottom) to generate three different speech parameters, i.e., "cv_mcep", "diff", and "diff_anasyn", to be fed into the waveform-processing module.

2.3. Conversion model with cascaded DMDN (CascDMDN)

To develop a more flexible spectral parameter model, the NU VC system employs a cascading structure of the DNN and DMDN called the cascaded DMDN (CascDMDN). In CascDMDN, two sets of hidden layers are used, where the first set is used to estimate the pdf of the source spectral parameters and the second one is used to estimate a mixture of pdfs of the target parameters. Therefore, the conditional pdf of the target spectral feature vector is defined as follows:

$$P(\boldsymbol{Y}_t|\boldsymbol{X}_t,\boldsymbol{\lambda}) \simeq P_s(\hat{\boldsymbol{X}}_t|\boldsymbol{X}_t,\boldsymbol{\lambda}_1)P_m(\boldsymbol{Y}_t|\hat{\boldsymbol{X}}_t,\boldsymbol{\lambda}_2), \quad (13)$$

where the parameters of the first set are denoted as $\boldsymbol{\lambda}_1$ and those of the second one are denoted as $\boldsymbol{\lambda}_2$. The set of network parameters of CascDMDN is denoted as $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2\}$. In the above likelihood function, the first conditional pdf is similar to that of the DNN in Eq. (1), while the second one refers to the mixture output layer of the DMDN as in Eq. (5). The representation of CascDMDN is given by the right graphical model in Fig. 1.

In the training phase of CascDMDN, a set of updated network parameters $\hat{\boldsymbol{\lambda}}$ is estimated by backpropagating the following loss:

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} -\prod_{t=1}^{T} P_s(\boldsymbol{X}_t|\boldsymbol{X}_t,\boldsymbol{\lambda}_1)P(\boldsymbol{Y}_t|f_{\lambda_1}(\boldsymbol{X}_t),\boldsymbol{\lambda}_2). \quad (14)$$

Note that not only does the estimation of the source spectral feature vector in the first set give flexibility in the parameter inference, such as, for computing spectral differences between the estimated target and source spectral parameters as shown in the upper part of Fig. 2, but it also provides an additional regularization term in model training.

Then, in the conversion phase, given a source spectral feature vector sequence $\boldsymbol{X}$, the trajectory of the target spectral parameters $\boldsymbol{y}$ is estimated in a similar manner to the MLPG of the DMDN in Eq. (9), where the mixture output layer is denoted as $f_{\lambda_2}(f_{\lambda_1}(\boldsymbol{X}_1))$. Following the structure of the network, the trajectory of the source spectral parameters $\hat{\boldsymbol{x}} = [\hat{\boldsymbol{x}}_1^\top, \hat{\boldsymbol{x}}_2^\top, \ldots, \hat{\boldsymbol{x}}_T^\top]^\top$ can be estimated as in the MLPG of the DNN in Eq. (4). In addition, the global variance (GV) [16] postfilter is applied to the converted spectral sequence to alleviate oversmoothed structures.

3. Waveform-processing module

The NU VC system uses a WaveNet-based vocoder [17, 18, 19] to model the waveform of the target speaker and generate the converted waveform using estimated speech features. Several flows are used in producing the estimated spectral features, where the direct waveform modification [2] method is employed. In addition, a selection procedure is performed to obtain the best waveform generation flow in an utterance-wise manner.

3.1. Analysis-synthesis with direct waveform modification

It is well known that vocoder-based waveform generation usually causes quality degradation in the generated speech owing to the difficulty of modeling source excitation signals. To avoid this issue, the direct waveform modification (DiffVC) method [2] has been proposed to directly filter an input waveform according to spectral differences between the target waveform and the input waveform. However, because the excitation features are not converted, it is difficult to convert speaker characteristics with a large difference in prosody characteristics, such as in a cross-gender conversion. Here, we describe an analysis-synthesis method to obtain refined spectral parameters that is based on the DiffVC method while making it possible to perform F0 conversion within the analysis-synthesis flow, as shown by the bottom flow in Fig. 2.

The analysis-synthesis procedure produces three different types of speech features to be fed into the WaveNet vocoder. The first one is called the "cv_mcep" set, which consists of the GV-postfiltered estimated target spectral parameters ("GV-PF-ed mcep") and the input band-aperiodicity features. Following this, the input waveform is then directly filtered ("diff-waveform") according to the spectral differences between GV-

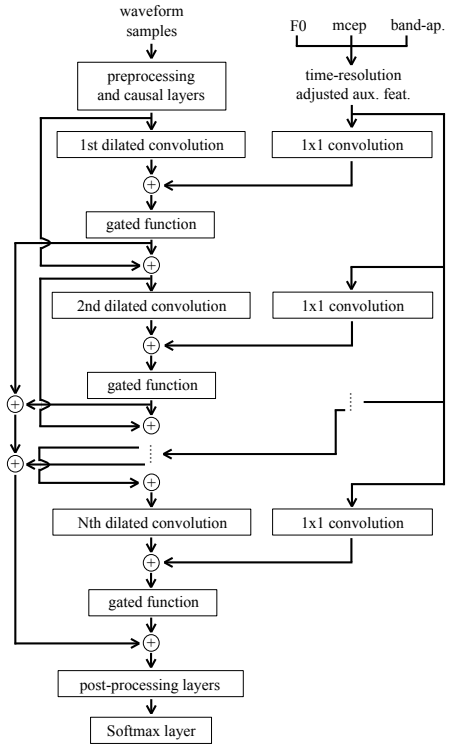

Figure 3: Architecture of the WaveNet vocoder having stacked residual blocks consisting of dilated convolutional layers to directly model waveform samples.

PF-ed mcep and the input spectral parameters. Then, by analyzing diff-waveform according to the original F0 values, the second feature set, called the "diff" set, which consists of the modified spectral and band-aperiodicity features, is obtained. Next, after performing the conversion of the F0 values, the F0-modified diff-waveform ("diff-conv-F0 waveform") is synthesized with a vocoder by using diff parameter set. Finally, the third set of parameters, called "diff_anasyn", is obtained by analyzing diff-conv-F0 waveform according to the converted F0 values. Note that the estimated target spectral parameters are generated in accordance with Section 2, while the converted F0 values are used by all three types of speech parameter set. Furthermore, we also investigated the use of the estimated input spectral parameters to compute the spectral differences ("diff-mcep"), as shown in the upper part of Fig. 2, although it did not yield significant improvements.

3.2. WaveNet-based waveform modeling and generation

The WaveNet vocoder [21], as illustrated in Fig. 3, is trained using the conditional pdf of each waveform sample with respect to their previous samples and possible auxiliary speech features. The likelihood function of a sequence of waveform samples $\boldsymbol{s} = [s_1, s_2, \ldots, s_T]^\top$ given a sequence of auxiliary features $\boldsymbol{h} = [\boldsymbol{h}_1^\top, \boldsymbol{h}_2^\top, \ldots, \boldsymbol{h}_T^\top]^\top$ is defined as follows:

$$P(\boldsymbol{s}|\boldsymbol{h}) = \prod_{t=1}^{T} P(s_t|s_1, s_2, \ldots, s_{t-1}, \boldsymbol{h}_t). \quad (15)$$

In modeling the speech waveform, WaveNet uses a stack of dilated convolutional layers, which makes it possible to exponentially increase the number of receptive fields on the waveform samples efficiently. In each step of the layer, there is a residual block consisting of a 2×1 dilated causal convolution with a gated activation function, and two 1×1 convolutions connected to either the next residual block or the skip connection. All skip connection outputs from the residual blocks are summed then fed to a post-processing layer. The gated activation function in a residual block including the auxiliary features is defined as follows:

$$\tanh\left(W_{f,k} * \boldsymbol{s} + V_{f,k} * \boldsymbol{h}'\right) \odot \sigma(W_{g,k} * \boldsymbol{s} + V_{g,k} * \boldsymbol{h}'), \quad (16)$$

where $W * \boldsymbol{s}$ denotes a dilated causal convolution, $V * \boldsymbol{h}'$ denotes a 1×1 convolution, σ denotes a sigmoid activation function, k is the layer index, f and g denote "filter" and "gate", respectively, and $\boldsymbol{h}'$ denotes transformed auxiliary features with the same time resolution as the speech waveform. The trainable convolution filters are denoted as W and V.

The NU VC system uses F0 information, mel-cepstrum parameters, and band-aperiodicity features as the auxiliary features for the WaveNet vocoder, where their time resolution is adjusted to that of the waveform by simply copying the values at respective frames. The F0 information includes unvoiced/voiced (U/V) binary decision and interpolated continuous F0 values. In the training of the model, these auxiliary features are extracted from the speech waveform. In the waveform generation phase, these parameters are obtained by one of the three different flows described in Section 3.1. Finally, the converted waveform is generated sample by sample.

3.3. Flow selection with an automatic waveform checker

The WaveNet-based vocoder is capable of generating much more natural-sounding waveforms than a conventional vocoder. However, sometimes the converted waveform generated by the WaveNet vocoder incorporates collapsed segments. This is most likely caused by the mismatch between the converted auxiliary features and the original features used in training the model. To recap, three different auxiliary features, described in Section 3.1, are considered in the WaveNet-based vocoder, i.e., the cv_mcep set, diff set, and diff_anasyn set, as also shown in Fig. 2. To select the best flow in the waveform generation, the NU VC system employs an automatic waveform checker to perform utterance-based selection.

To select the best waveform generation flow for each utterance, a power-based detector is employed to automatically detect collapsed segments in WaveNet-generated waveforms. From the spectrum of a generated waveform, frame-based summation is performed using the power spectrum of all frequency bins $\boldsymbol{P}$ and that of the Nyquist frequency components $\boldsymbol{L}$. Let $\boldsymbol{P}^{(W)} = [P_1^{(W)}, \ldots, P_T^{(W)}]$ and $\boldsymbol{P}^{(C)} = [P_1^{(C)}, \ldots, P_T^{(C)}]$ denote the power summation sequence from all frequency bins of a WaveNet-generated waveform and that of a conventional vocoder, respectively. The power summation sequences from the Nyquist frequency components are respectively denoted as $\boldsymbol{L}^{(W)} = [L_1^{(W)}, \ldots, L_T^{(W)}]$ and $\boldsymbol{L}^{(C)} = [L_1^{(C)}, \ldots, L_T^{(C)}]$. In the detection, the differences in the maximum power between the WaveNet-generated waveform and that generated the conventional vocoder are computed as follows:

$$\Delta\boldsymbol{P} = \max(\boldsymbol{P}^{(W)}) - \max(\boldsymbol{P}^{(C)}) \quad (17)$$

$$\Delta\boldsymbol{L} = \max(\boldsymbol{L}^{(W)}) - \max(\boldsymbol{L}^{(C)}). \quad (18)$$

The system selects the best waveform generation flow through the comparison of $\Delta\boldsymbol{P}$ and $\Delta\boldsymbol{L}$ with an empirical threshold, where both values will be higher than the threshold for a low-quality waveform.

4. Experiments and results

4.1. Experimental conditions

The speech database for the HUB task of the VCC 2018 consisted of four source speakers and four target speakers, which had two female and two male speakers for the source and another two female and two male speakers for the target. In the training set, each speaker uttered the same set of 81 English sentences, whereas the evaluation set consisted only of the four source speakers uttering another set of 35 sentences. The speech signal sampling rate was 22,050 Hz. The WORLD [23, 24] package was used in speech analysis. From a speech signal, 35-dimensional mel-cepstrum parameters including the 0th power coefficient, F0 values, and 513-dimensional aperiodicity features, which were coded into two-band aperiodicity parameters, were used. The frame shift was set to 5 ms.

Following the spectral parameter conversion module described in Section 2, the DNN used four hidden layers. On the other hand, the DMDN used a total of three hidden layers and 16 mixture components. CascDMDN, which is a combination of these two structures, used one hidden layer for estimating source spectral parameters and four hidden layers with 16 mixture components for estimating target spectral parameters. ReLU activation function was used for the hidden units. For every model, the learning rate was set to 0.0006, the weights were initialized with the Xavier [25] method, the initial biases were set to zero, the Adam [26] optimization was employed, and an utterance-size batch was used.

The NU VC system used the WaveNet-based vocoder described in Section 3.2 for waveform modeling and generation. The hyperparameters of the WaveNet vocoder are as follows: the learning rate was set to 0.001 with a decay factor of 0.5 per 50,000 iteration steps, 20,000 batch-size samples were used with a total of 200,000 iteration steps, the number of residual blocks was 30, the dilation sequence was $1, 2, 4, \ldots, 512$ with three repetitions, the number of channels for residual blocks and dilated causal convolution was 512, the number of channels for skip connection was 256, and the Adam [26] algorithm was used for optimization. To train the WaveNet model, a speaker-independent (SI) network was first trained by using all the data of eight speakers in the HUB task plus the data of four speakers from the SPOKE task, i.e., with another 81 different sets of utterances, and the data of two speakers from the ARCTIC database, i.e., "rms" and "slt", with each having 1132 utterances. The SI-WaveNet model was then fine-tuned by updating only the output layers using the data of each of the four target speakers, which resulted in four WaveNet models.

In the waveform generation phase, the three different auxiliary features described in Section 3.1 were considered, i.e., cv_mcep, diff, and diff_anasyn, as shown in Fig. 2. The list of priorities was made heuristically, with the diff_anasyn set at the top followed by the diff set. As described in Section 3.3, to avoid collapsed segments in the WaveNet-generated waveforms, the NU VC system used a flow selection procedure to rule out waveforms with low quality.

The results of using mel-cepstral distortion to evaluate the spectral conversion module are given in the objective evaluation results. An internal subjective evaluation was conducted to assess the performance of the NU VC system with the provided baseline system, i.e., "sprocket" [27], where the results are given in the internal subjective evaluation section. Finally, the last three sections describe the official results of the subjective evaluation in VCC 2018. Note that the results for the SPOKE task (nonparallel data) are presented in [28].

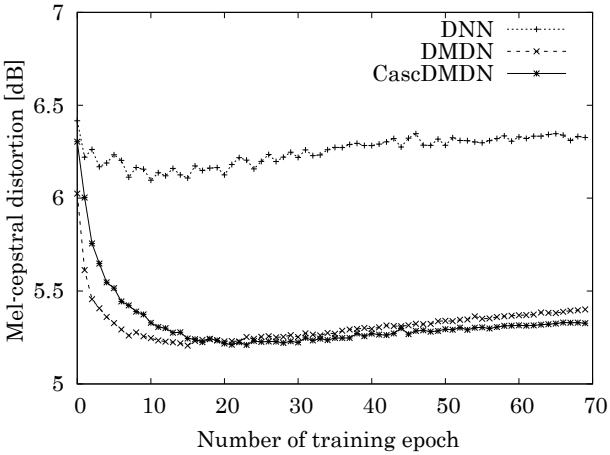

Figure 4: Plot of mel-cepstral distortions for DNN, DMDN, and CascDMDN, measured with the first 10 utterances excluded from the training dataset.

4.2. Objective evaluation

To compare the several deep learning models presented in Section 2, an evaluation of mel-cepstral distortion was performed for the DNN, DMDN, and CascDMDN for the spectral parameter estimation. In this objective evaluation, the first 10 utterances from the training dataset were excluded while training the models. Then, they were used to compute the mel-cepstral distortion between the extracted target mel-cepstrum parameters and the estimated values as follows:

$$\text{Mel-CD[dB]}_t = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{34}(y_t(d) - \hat{y}_t(d))^2}, \qquad (19)$$

where $y_t(d)$ and $\hat{y}_t(d)$ denote the dth dimension of the extracted mel-cepstrum parameters and that of the estimated values at frame t, respectively.

The trends of the mel-cepstral distortion averaged over all 16 speaker pairs during 70 training epochs are shown in Fig. 4. It can be observed that CascDMDN is capable of providing much more stable distortion than the conventional DNN and slightly more stable distortion than the conventional DMDN. The flexibility of the CascDMDN structure in providing access to estimated source spectral parameters makes a good choice for the spectral conversion model in the NU VC system. The overfitting condition observed in this objective evaluation served as a reference in training the final model using all training data.

4.3. Internal subjective evaluation

In the internal subjective evaluation, two preference tests (naturalness and speaker similarity) were conducted to compare the performance of the NU VC system with that of the baseline system, i.e., sprocket [27]. All 16 speaker pair models for the four source and four target speakers were used in the evaluation. The total number of available evaluation utterances was 35. The total number of listeners was eight and none of them were native English speakers. In the naturalness test, two audio samples, one each for the NU and the baseline system, of the same utterance were presented to a listener in a random order. Then, the listener was asked to select the audio preference according to naturalness. Meanwhile, in the preference test, in addition to two generated audio samples, two original audio samples of the corresponding target speaker randomly taken from the training

Table 1: Result of naturalness preference test in the internal subjective evaluation of the NU VC system and the baseline (sprocket) for same-gender and cross-gender conversions.

Naturalness	Same-gender	Cross-gender
Baseline	$68\% \pm 7\%$	$42\% \pm 7\%$
NU	$32\% \pm 7\%$	$58\% \pm 7\%$

Table 2: Result of speaker identity preference test in the internal subjective evaluation of the NU VC system and the baseline (sprocket) for same-gender and cross-gender conversions.

Spk. Identity	Same-gender	Cross-gender
Baseline	$43\% \pm 7\%$	$45\% \pm 8\%$
NU	$57\% \pm 7\%$	$55\% \pm 8\%$

dataset were presented. The listener was then asked to select their preference based on the similarity to the target speaker characteristic. From the 35 evaluation utterances, three were randomly taken for each speaker pair in each preference test, resulting in a total of 92 audio samples for each listener.

The results of the internal subjective evaluation are summarized in Tables 1 and 2. It can be observed that the baseline system achieves a significantly higher preference score in terms of naturalness for the same-gender conversions, with a score of 68%, compared with 32% for the NU system. However, the NU system yields a higher naturalness preference score for the cross-gender conversions, with a score of 58%, compared with 42% for the baseline. On the other hand, in the preference test for speaker similarity, the NU system achieves higher preference scores for both same- and cross-gender conversions, with scores of 57% and 55%, compared with 43% and 45% for the baseline, respectively. This result is reasonable because the baseline, i.e., sprocket, uses vocoder-free waveform generation for same-gender conversions and vocoder-based generation for the cross-gender conversions. This implies that the use of a WaveNet-based vocoder can improve the generated waveform quality compared with that obtained using the conventional vocoder and gives much higher accuracy than both the conventional vocoder and the vocoder-free system, i.e., direct waveform modification.

4.4. Official subjective evaluation

In VCC 2018, to compare the performance of the submitted systems, an official subjective evaluation was conducted, which consists of a mean opinion score (MOS) test on the speech quality and a speaker similarity test. In the MOS test, each listener was given stimuli of audio samples and asked to evaluate the naturalness of the speech sounds using a five-point scale (1: Completely unnatural; 2: Mostly unnatural; 3: Equally natural and unnatural; 4: Mostly natural; 5: Completely natural). In the speaker similarity test, each listener was given a pair of audio samples as stimuli and asked to judge whether they were produced by the same speaker. Their confidence in the decision was given on a four-point scale (1: Same, absolutely sure; 2: Same, not sure; 3: Different, not sure; 4: Different, absolutely sure). The total number of listeners was 106 (49 female, 57 male).

The results of the official objective evaluation are summarized in Fig. 5. The results show the average MOS in terms of speech quality, plotted on the x-axis, for every submitted system, including the baseline (sprocket), as well as the original source and target speakers. The similarity scores (in %), computed by adding up the two confidence scores in each binary similarity decision, are plotted on the y-axis. The NU VC system, denoted as N17, achieves an MOS of 3.44 for speech quality, compared with 3.57 for the baseline, 4.15 for the top system, i.e., N10, and also 3.44 for the closest system, i.e., N08. On the other hand, for the speaker similarity test, the NU VC system achieves a similarity score of 85%, outperforming the baseline (68%) and all other systems except system N10, which has slightly higher accuracy (86%). Overall, the NU VC system was placed as the runner-up behind the top system, N10. Details of the official subjective evaluation results are given in the following sections.

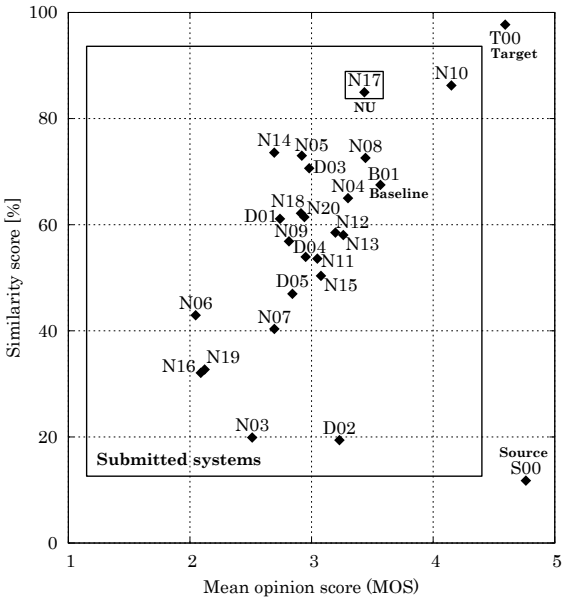

Figure 5: Scatter plot of mean opinion score (MOS) for speech quality and speaker similarity score for the submitted systems along with baseline (sprocket) [B01], source [S00], and target [T00] speech. The NU VC system is N17.

4.5. Detailed results for speech quality

The detailed official results of the MOS for speech quality for all systems including the baseline, are given in Figs. 6 and 7. The results for same-gender conversions, which consist of female-to-female (F-F) and male-to-male (M-M) conversions, are shown in the Fig. 6, whereas, those for the cross-gender conversions, i.e., female-tomale (F-M) and male-to-female (M-F), are shown in Fig. 7.

The NU VC system, denoted as N17, achieves an MOS of 3.24 for the cross-gender conversions and 3.63 for the same-gender conversions, which place the system in the fourth and the third places, respectively. The MOSs for each gender conversion are 3.89 for F-F, 3.38 for M-M, and 3.24 for both F-M and M-F. Compared with the baseline, it is expected that the NU system will perform better in cross-gender conversions because sprocket uses the vocoder-based method in these conversions. The MOSs for the baseline system are 4.10, 3.88, 3.31, and 3.00 for the above conversions, respectively. However, the NU system is outperformed by the top system, i.e., N10, which achieves an MOS of over 4.10 for every gender-type conversion. Overall, compared with the other submitted systems, the

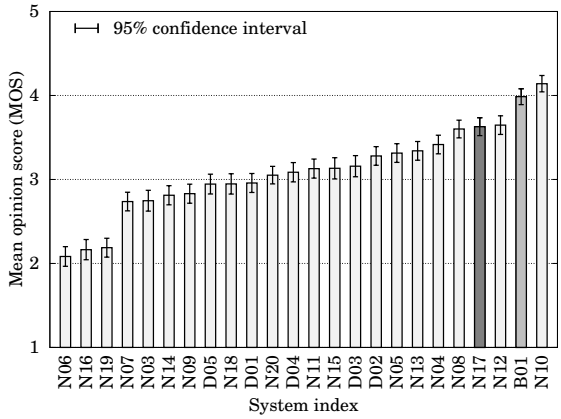

Figure 6: MOS results for speech quality for same-gender conversions, i.e., female-to-female (F-F) and male-to-male (M-M) conversions. The NU VC system is N17.

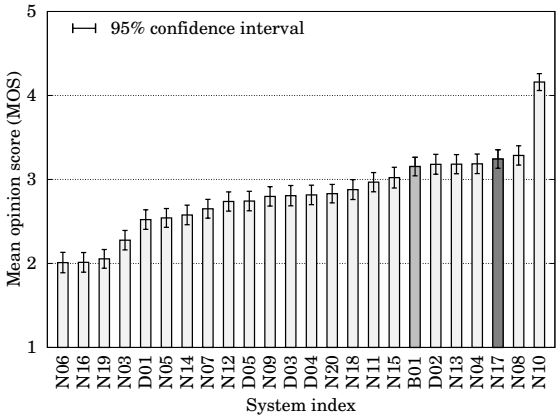

Figure 7: MOS results for speech quality for cross-gender conversions, i.e., female-to-male (F-M) and male-to-female (M-F) conversions. The NU VC system is N17.

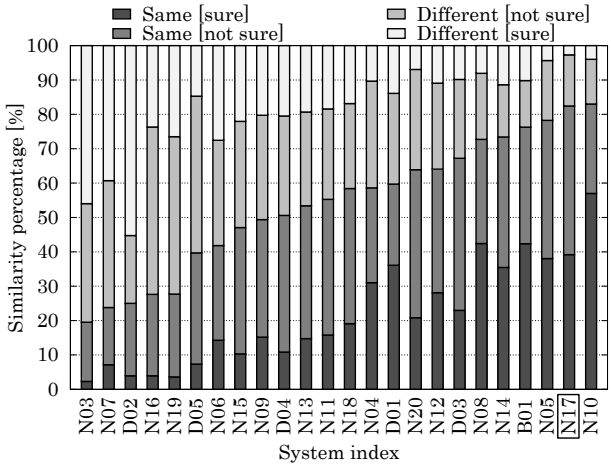

Figure 8: Similarity percentage results for same-gender conversions, i.e., F-F and M-M conversions with two confidence levels in the binary decision. The NU VC system is N17.

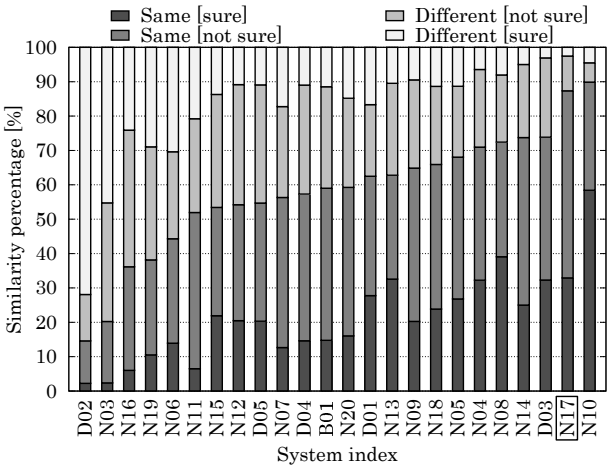

Figure 9: Similarity percentage results for cross-gender conversions, i.e., F-M and M-F conversions with two confidence levels in the binary decision. The NU VC system is N17.

NU system yields a good performance with an average MOS of 3.44 over all speaker pairs, the same as the system N08, slightly below those of the baseline (3.57) and far behind system N10 (4.15).

4.6. Detailed results for speaker similarity

The official results for the speaker similarity evaluation are shown in Figs. 8 and 9. The results for same-gender conversions, i.e., F-F and M-M, are shown in Fig. 8, whereas the result for cross-gender conversions, i.e., F-M and M-F, are shown in Fig. 9. These figures show the percentage of speaker similarity decisions, i.e., "same" or "different", each with two confidence levels, i.e., "sure" and "not sure". To measure the final similarity score, the percentage scores of "same" ("sure") and "same" ("not sure") are added together.

The NU VC system (N17) has a total similarity score of 82% ("same" decisions), i.e., 18% "different" decisions for the same-gender conversions, and a similarity score of 87% for the cross-gender conversions. The details for each gender conversion are as follows: similarity scores of 82% for F-F conversion, 83% for M-M conversion, 93% for F-M conversion, and 76% for M-F conversion. In this speaker similarity evaluation, the NU system outperforms the baseline, as shown in both Figs. 8 and 9, where the baseline has the following similarity scores in the same order: 84%, 53%, 59%, and 60%. Compared with the top system, i.e., N10, the NU VC system yields similar results: where our system has a slightly better scores in F-F conversion (81% for N10) and in F-M conversion (91% for N10), a lower score in M-F conversion (85% for N10), and a much lower score in M-M conversion (94% for N10). Overall, the NU VC system yields a very good performance with an average similarity score of 85% over all speaker pairs, outperforming the baseline (68%) and all of the systems (the closest are N14 with 74%, and both N05 and N08 with 73%) except for N10, which has a slightly higher score of 86%.

5. Conclusion

In this paper, the NU (Nagoya University) voice conversion (VC) system developed for the HUB task of the Voice Conversion Challenge 2018 has been presented. The NU VC system adopts a deep learning architecture to develop a spectral parameter conversion model by combining a deep neural network (DNN) and deep mixture density network (DMDN) to form

a cascaded DMDN (CascDMDN). In the waveform modeling and generation, the NU VC system employs a WaveNet-based vocoder. The auxiliary features fed into the WaveNet system are chosen from several analysis-synthesis flows using a model selection procedure in an utterance-wise manner. The results of the challenge put the NU VC system in the second place with an average mean opinion score (MOS) of 3.44 for speech quality and a similarity score of 85% for speaker identity.

6. Acknowledgements

This work was partly supported by JST, PRESTO Grant Number JPMJPR1657, and JSPS KAKENHI Grant Number JP17H06101.

7. References

[1] F. Villavicencio and J. Bonada, "Applying voice conversion to concatenative singing-voice synthesis," in *Proc. INTERSPEECH*, Sep. 2010, pp. 2162–2165.

[2] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in *Proc. INTERSPEECH*, Singapore, Sep. 2014, pp. 2514–2518.

[3] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 9, pp. 2505–2517, 2012.

[4] A. B. Kain, J.-P. Hosom, X. Niu, J. P. van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Commun.*, vol. 49, no. 9, pp. 743–759, 2007.

[5] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A hybrid approach to electrolaryngeal speech enhancement based on spectral subtraction and statistical voice conversion," in *Proc. INTERSPEECH*, Lyon, France, Sep. 2013, pp. 3067–3071.

[6] P. L. Tobing, K. Kobayashi, and T. Toda, "Articulatory controllable speech modification based on statistical inversion and production mappings," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 12, pp. 2337–2350, 2017.

[7] T. Toda, L.-H. Chen, F. Villavicencio, Z. Wu, and J. Yamagishi, "The Voice Conversion Challenge 2016," in *Proc. INTERSPEECH*, San Francisco, USA, Sep. 2016, pp. 1632–1636.

[8] 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," Submitted to Odyssey 2018.

[9] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *J. Acoust. Soc. of Jpn. (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[10] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, 1998.

[11] 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 Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 1859–1872, 2014.

[12] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv preprint arXiv:1711.11293*, 2017.

[13] D. T. Chappell and J. H. Hansen, "Speaker-specific pitch contour modeling and modification," in *Proc. ICASSP*, Seattle, USA, May 1998, pp. 885–888.

[14] Z.-Z. Wu, T. Kinnunen, E. S. Chng, and H. Li, "Text-independent F0 transformation with non-parallel data for voice conversion," in *Proc. INTERSPEECH*, Makuhari, Japan, Sep. 2010, pp. 1732–1735.

[15] L.-H. Chen, L.-J. Liu, Z.-H. Ling, Y. Jiang, and L.-R. Dai, "The USTC system for Voice Conversion Challenge 2016: Neural network based approaches for spectrum, aperiodicity and F0 conversion," in *Proc. INTERSPEECH*, San Francisco, USA, Sep. 2016, pp. 1642–1646.

[16] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.

[17] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent wavenet vocoder," in *Proc. INTERSPEECH*, Aug. 2017, pp. 1118–1122.

[18] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for wavenet vocoder," in *Proc. ASRU*, Dec. 2017.

[19] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with wavenet-based waveform generation," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 1138–1142.

[20] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 3844–3848.

[21] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[22] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP*, Detroit, USA, May 1995, pp. 660–663.

[23] 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. 99, no. 7, pp. 1877–1884, 2016.

[24] C.-C. Hsu. Python-wrapper-for-world-vocoder. [Online]. Available: https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder

[25] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, Sardinia, Italy, May 2010, pp. 249–256.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[27] K. Kobayashi and T. Toda, "sprocket: Open-Source Voice Conversion Software," Submitted to Odyssey 2018.

[28] Y.-C. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, "The NU non-parallel voice conversion system for the Voice Conversion Challenge 2018," Submitted to Odyssey 2018.