THE NU NON-PARALLEL VOICE CONVERSION SYSTEM FOR THE VOICE CONVERSION CHALLENGE 2018

Yi-Chiao Wu[1], *Patrick Lumban Tobing*[1], *Tomoki Hayashi*[2], *Kazuhiro Kobayashi*[3], *Tomoki Toda*[3]

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

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

ABSTRACT

This paper presents the non-parallel voice conversion (VC) system developed at Nagoya University (NU) for the SPOKE task of the Voice Conversion Challenge 2018 (VCC2018). The goal of this task is to develop VC systems without the requirement of parallel training data. The key idea of our system is to use a text-to-speech (TTS) voice as a reference voice, making it possible to create two parallel training datasets between the source and TTS voices and between the TTS and target voices. Using these datasets, a cascade VC system is developed to convert the source voice into the target voice via the TTS voice as a reference. Furthermore, we also propose a system selection framework to avoid generating collapsed speech waveforms, which are often observed when using less accurately converted speech features in the WaveNet vocoder. In VCC2018, our system achieved the second best score for similarity (around 70%) and an above-average score for naturalness (a mean opinion score around 3.0) among the submitted systems.

1. INTRODUCTION

Voice conversion (VC) is a generation problem of machine learning. Given an input speech and information of a specific target, a machine should be able to generate the target speech corresponding to the linguistic contents of the given input speech. A typical VC application is speaker conversion to convert the speaker identity of a source speaker's voice to that of a specified target speaker's voice. For simplicity, we use the term VC in this paper to denote speaker conversion.

Numerous VC approaches have been proposed, such as Gaussian mixture model (GMM) [1, 2], frequency warping [3, 4], deep neural network (DNN) [5-7], and exemplar-based approaches [8-10]. The traditional VC framework usually requires a parallel speech data of a specific source and target speaker pair to construct a mapping function. However, the requirement of a parallel corpus causes limitations in practical applications. Therefore, many attempts have been made to build a flexible VC system without the requirement of a parallel corpus. For example, the INCA [11] algorithm is a representative method of aligning non-parallel data, which iteratively trains a VC function based on the nearest-neighbors alignment between intermediate converted and target voices, then generating an updated intermediate voice closer to the target voice to obtain better frame wise alignment in each iteration. Another approach is to separate the speaker information and context information using AutoEncoder [12, 13]. Furthermore, it also has been proven effective to use a speech recognizer to retrieve context information to map source and target acoustic features [14, 15]. These nonparallel VC techniques have great potential for developing more flexible VC systems handling arbitrary speakers, such as many-to-one, one-to-many, and many-to-many frameworks [16, 17]. One of the practical approaches defining the mapping between an arbitrary speaker pair is to use an intermediate speaker as a reference [17-19].

In this paper, we focus on the SPOKE task of the Voice Conversion Challenge 2018 (VCC2018) [20], for which a non-parallel corpus and corresponding transcripts of source and target speakers are provided. The design of the proposed Nagoya University (NU) non-parallel VC system is based on the assumption that although text-to-speech (TTS) output is not natural, it still has sufficient acoustic components as natural speech to be used as a reference. Therefore, we can apply the TTS output as the middle-layer reference to derive the mapping relationship between the non-parallel source and target utterances. In addition, we apply the WaveNet vocoder [21-23] for speech generation to generate natural speech waveforms. However, the mismatch of training features and converted features sometimes causes the WaveNet vocoder to suffer from collapsed speech waveforms [22]. Therefore, we also introduce a system selection technique to select generated utterances without the problem of collapsed speech waveforms.

The rest of this paper is organized as follows. The basic NU parallel VC system is described in Section 2. The proposed NU non-parallel VC system is introduced in Section 3. The experimental results are presented in Section 4. Finally, the conclusion is given in Section 5.

2. BASIC STRUCTURE OF THE NU VC SYSTEM

In this section, we describe the basic NU parallel VC system. Figure 1 shows the overall system, which includes DNN-based spectral conversion, analysis-synthesis framework with direct waveform modification (DIFFVC) [24] for excitation signal transformation, linear fundamental frequency (F_0) conversion, and speech generation using WaveNet [25]. First, given an input speech waveform, speech parameters, including spectral features, F_0, and aperiodicity features (*ap*), are extracted using WORLD [26, 27] analysis. Then, a frame-based spectral conversion procedure is performed using DNN, while the F_0 is linearly transformed using the statistics of the training data. After obtaining the converted F_0 and spectral features, an

10.21437/Odyssey.2018-30

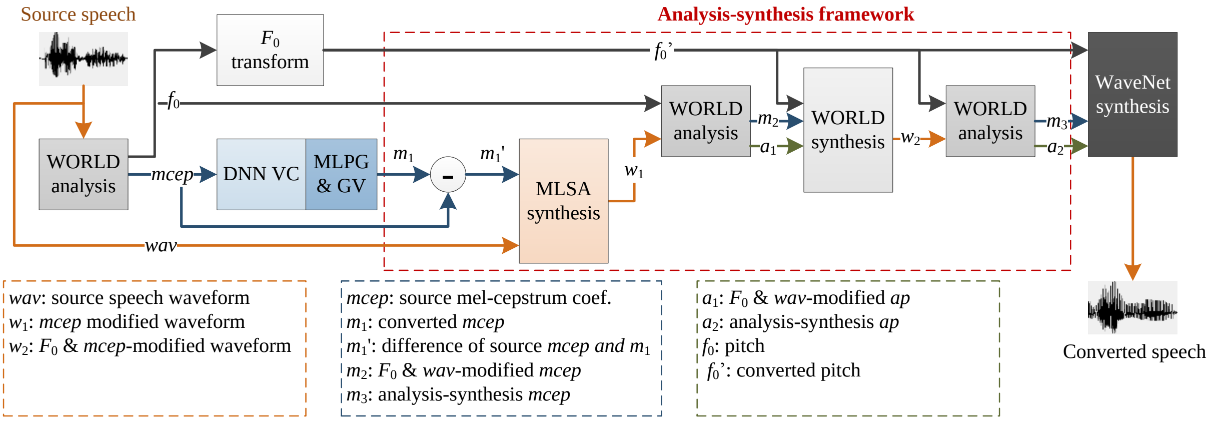
Source speech
Analysis-synthesis framework
F_0 transform
f_0'
f_0
WORLD analysis
mcep
DNN VC
MLPG & GV
m_1
-
m_1'
MLSA synthesis
w_1
wav
WORLD analysis
m_2
a_1
WORLD synthesis
w_2
WORLD analysis
m_3
a_2
WaveNet synthesis
Converted speech
wav: source speech waveform
w_1: *mcep* modified waveform
w_2: F_0 & *mcep*-modified waveform
mcep: source mel-cepstrum coef.
m_1: converted *mcep*
m_1': difference of source *mcep and* m_1
m_2: F_0 & *wav*-modified *mcep*
m_3: analysis-synthesis *mcep*
a_1: F_0 & *wav*-modified *ap*
a_2: analysis-synthesis *ap*
f_0: pitch
f_0': converted pitch

Figure 1. *Basic NU parallel VC system based on DNN and WaveNet vocoder*

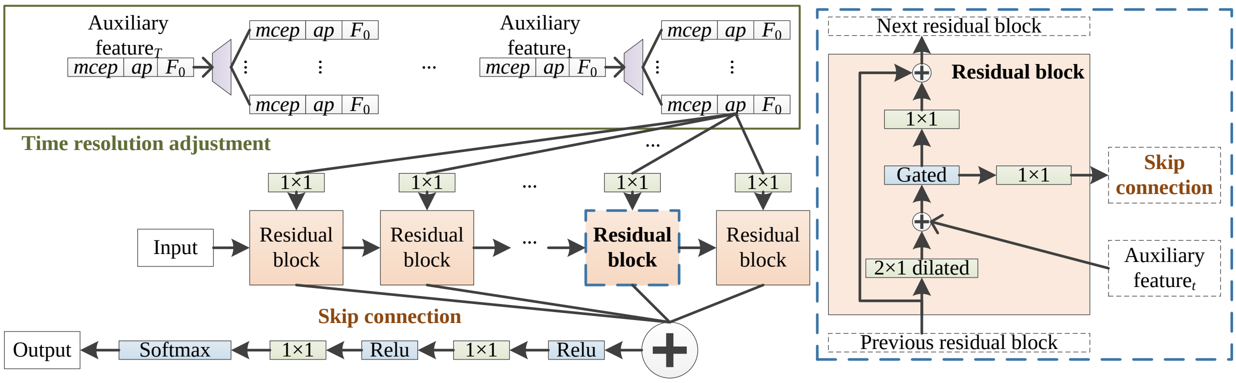

Figure 2. *Conditional WaveNet vocoder architecture*

analysis-synthesis framework based on direct waveform modification is performed to further modify the spectral and excitation signals. Finally, given the coded *ap*, converted F_0, and converted spectral features, the converted speech waveform is generated using the WaveNet vocoder [21-23].

2.1. Spectral feature mapping based on DNN

DNN-based spectral conversion [28, 29] consists of training and conversion stages. Given the source static-dynamic feature vector $\mathbf{X}_t = \left[\mathbf{x}_t^{\mathrm{T}}, \Delta\mathbf{x}_t^{\mathrm{T}}\right]^{\mathrm{T}}$, the conditional probability density function of the target static-dynamic feature vector $\mathbf{Y}_t = \left[\mathbf{y}_t^{\mathrm{T}}, \Delta\mathbf{y}_t^{\mathrm{T}}\right]^{\mathrm{T}}$ at frame t is given by

$$P\left(\mathbf{Y}_t \mid \mathbf{X}_t, \mathbf{D}, \boldsymbol{\lambda}\right) = \mathbb{N}\left(\mathbf{Y}_t; \mathrm{f}_{\boldsymbol{\lambda}}\left(\mathbf{X}_t\right), \mathbf{D}\right), \tag{1}$$

where $\mathbb{N}(\cdot)$ denotes a Gaussian distribution, $\mathrm{f}_{\boldsymbol{\lambda}}(\cdot)$ is a nonlinear conversion function from the DNN, $\boldsymbol{\lambda}$ represents the DNN parameters, and $\mathbf{D}$ is the diagonal covariance matrix of the training data. In the training stage, the estimation of the updated DNN parameters $\hat{\boldsymbol{\lambda}}$ is performed as follows:

$$\begin{aligned}\hat{\boldsymbol{\lambda}} &= \underset{\boldsymbol{\lambda}}{\operatorname{argmax}} \sum_{t=1}^{T} \log P\left(\mathbf{Y}_t \mid \mathbf{X}_t, \mathbf{D}, \boldsymbol{\lambda}\right) \\ &= \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \frac{1}{2} \sum_{t=1}^{T} \left(\mathbf{Y}_t - \mathrm{f}_{\boldsymbol{\lambda}}\left(\mathbf{X}_t\right)\right)^{\mathrm{T}} \mathbf{D}^{-1} \left(\mathbf{Y}_t - \mathrm{f}_{\boldsymbol{\lambda}}\left(\mathbf{X}_t\right)\right).\end{aligned} \tag{2}$$

In the conversion stage, given the DNN output, the trajectory of the target feature vector is generated by maximum likelihood parameter generation (MLPG) [30]. Furthermore, to alleviate the over-smoothing effect caused by the averaging factor in the model, a global variance (GV) [2] post-filter is applied to the converted trajectory.

2.2. Analysis-synthesis framework

The main propose of analysis-synthesis framework is to extract much matched excitation and spectral features for the WaveNet vocoder. To be more specific, the input waveform is first filtered according to the difference between converted and source spectra using DIFFVC in [24] to modify the spectral envelopes. Secondly, we modify the excitation signals after obtaining the spectral modified waveform. Specifically, the modified waveform is analyzed to extract the modified *ap* and the spectral features, and a frame-based linear transformation of the logarithmic F_0 is performed by lining up the pitch difference between the source and target speakers. After that, the converted waveform with converted spectral and excitation signals can be generated using the converted F_0, the modified *ap* and spectral features, and the WaveNet or WORLD vocoder. Moreover, according to our informal listening tests, the quality of speech obtained using the WaveNet vocoder as a post-filter is slightly higher than that of speech directly synthesized by the WaveNet vocoder. Therefore, on the basis of the converted F_0, the final converted *ap* and spectral features are extracted from the converted waveform generated by WORLD. The final converted speech is generated using the WaveNet vocoder on the basis of the final converted *ap* and spectral features and the converted F_0.

2.3. WaveNet vocoder

To generate more natural-sounding speech, in our system, the conventional vocoder (ex: WORLD) is replaced by the state-of-the-art WaveNet vocoder [21-23] to generate the final converted waveform. WaveNet [25] is a deep autoregressive network capable of directly modeling a speech waveform sample-by-sample using the following conditional probability:

$$P(\mathbf{Y}\,|\,\mathbf{h}) = \prod_{n=1}^{N} P\left(y_n \,|\, y_{n-r},\ldots,y_{n-1},\mathbf{h}\right), \tag{3}$$

where n is the sample index, r is the size of the receptive field, y_n is the current audio sample, and $\mathbf{h}$ is the vector of the auxiliary features. In our system, the auxiliary features consist of the coded *ap*, the transformed F_0, and the converted spectral features.

Figure 2 shows the structure of the WaveNet vocoder, which consists of many residual blocks, each block containing a 2 × 1 convolution dilated causal convolution, a gated activation function, and 1 × 1 convolutions. The dilated causal convolution is a convolution with a skipping value filter, which enables the network to efficiently operate on a large receptive field. The gated activation function is formulated as

$$\tanh\left(V_{f,k}^{1} * \mathbf{Y} + V_{f,k}^{2} * a(\mathbf{h})\right) \odot \sigma\left(V_{g,k}^{1} * \mathbf{Y} + V_{g,k}^{2} * a(\mathbf{h})\right), \tag{4}$$

where V^1 and V^2 are trainable convolution filters, $*$ is the convolution operator, $\odot$ is an element wise multiplication operator, σ is a sigmoid function, k is the layer index, f and g represent the "filter" and "gate", respectively, and $a(\cdot)$ is the resolution adjustment function used to duplicate auxiliary features to match the resolution of input speech samples. Furthermore, the input waveforms are quantized to 8 bits based on μ-law encoding and the generated waveforms are restored by μ-law decoding.

3. NU NON-PARALLEL VC SYSTEM

Figure 3 illustrates the system architecture of the proposed NU non-parallel VC system. The main concept is that instead of directly aligning the non-parallel source and target features, we map the source features to the target features with the assistance of reference speech. In addition, we also propose a collapsed speech detection technique for system selection.

3.1. Non-parallel VC with reference speaker

In the non-parallel spectral feature conversion, we construct the cascade VC system, which includes an encoder model to map the source features to the reference features and a decoder model to map the reference features to the target features. Specifically, we generate the reference speech corresponding to each of the source and target speakers by TTS to develop the parallel corpora. Then, in the training stage, we develop encoder models to convert the source speakers into the TTS speaker, and decoder models to convert the TTS speaker into the target speakers. Specifically, the encoder and decoder models are DNN-based spectral conversion models. As shown in Fig. 3, the speech spectral features contain speaker-dependent components (speaker information) and speaker-independent components (linguistic information). Both the encoding process and the decoding process only change the speaker-dependent parts.

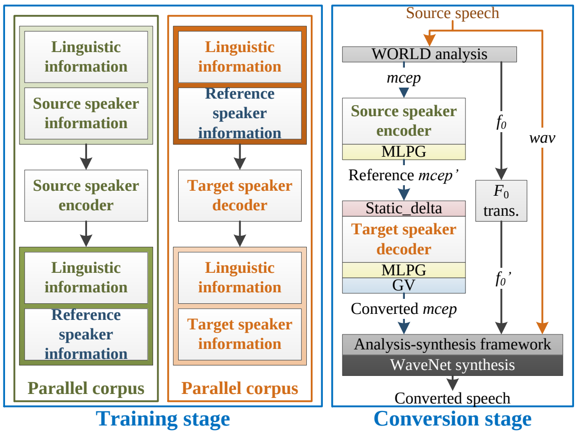

Figure 3. *NU non-parallel VC system*

In the spectral conversion stage, the system converts the source spectral features to the reference spectral features using the encoder, and then converts the reference spectral features to the target spectral features using the decoder. Moreover, although the GV [2] post-filter can improve the perceptual speech quality, it also enhances the prediction error. Therefore, the GV post-filter is only applied to the final output of the decoder.

3.2. System selection

Although the quality of speech generated from the WaveNet vocoder is usually higher than that generated from a conventional vocoder (ex: WORLD), the WaveNet vocoder is more sensitive to less accurately converted spectral features [22]. As shown in Fig. 4, for the same converted features, the WaveNet vocoder sometimes generates collapsed waveform segments, while the outputs of the conventional vocoder tend to be more stable. This problem is likely caused by the mismatch between the target training features and the online converted features. Therefore, a collapsed speech detection technique is introduced to evaluate the quality of each WaveNet generated utterance, and its results are used for the utterance-based system selection.

For the detection technique, we find that although the waveforms are different, the powers of the generated utterances from the WaveNet vocoder and the conventional vocoder are very similar. Therefore, the differences of maximum powers will become large when the utterance has a collapsed segment, particularly in the high-frequency band. In accordance with this observation, we then design a detection criterion based on the frame-based summations of the power spectrum (denoted as $\mathbf{P}$) and the power of Nyquist frequency components (denoted as $\mathbf{L}$). Given the power sequences $\mathbf{P}^{(W)} = \left[p_1^{(W)},\ldots,p_T^{(W)}\right]$ and $\mathbf{P}^{(C)} = \left[p_1^{(C)},\ldots,p_T^{(C)}\right]$, and the power sequences of Nyquist frequency $\mathbf{L}^{(W)} = \left[l_1^{(W)},\ldots,l_T^{(W)}\right]$ and $\mathbf{L}^{(C)} = \left[l_1^{(C)},\ldots,l_T^{(C)}\right]$, the detection measurements are defined as

$$\Delta\mathbf{P} = \max\left(\mathbf{P}^{(W)}\right) - \max\left(\mathbf{P}^{(C)}\right) \tag{5}$$

and

$$\Delta\mathbf{L} = \max\left(\mathbf{L}^{(W)}\right) - \max\left(\mathbf{L}^{(C)}\right), \tag{6}$$

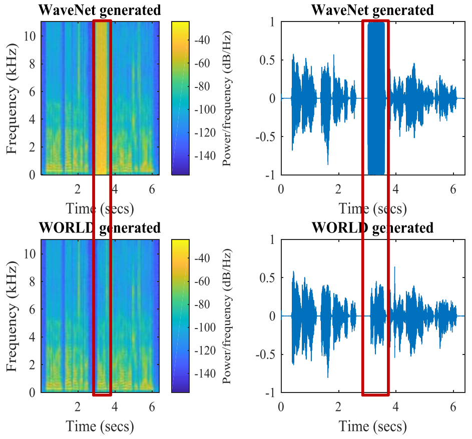

Figure 4. *For the same converted features, the WaveNet vocoder generated collapsed waveform (upper right) and spectrogram (upper left), and the WORLD vocoder generated normal waveform (bottom right) and spectrogram (bottom left) are shown.*

where W denotes the utterance generated from the WaveNet vocoder, C denotes the utterance generated from the conventional vocoder, and T is the number of frames. If both $\Delta\mathbf{P}$ and $\Delta\mathbf{L}$ are higher than an empirical threshold, the system selects the utterance from the conventional vocoder. Moreover, we use the differences between maximum powers instead of the frame-based power differences because of their stability.

4. EXPERIMENTAL EVALUATIONS

In this section, the internal objective evaluations and the external subjective tests carried out in VCC2018 are reported.

4.1. Experimental conditions

The evaluation corpus was an English speech dataset provided by the VCC2018 organizer. The corpus included two subsets, HUB and SPOKE. The HUB corpus consisted of four male speakers and four female speakers. Two males and two females were the source speakers, and the remaining four speakers were the target speakers for the parallel VC task (HUB task). Each speaker in the HUB set had 81 parallel utterances for training, and each source speaker had 35 parallel utterances for testing. On the other hand, the SPOKE corpus included another two male and two female speakers as the source speakers for the non-parallel VC task (SPOKE task). Each speaker in the SPOKE set also had 81 parallel utterances for training and 35 parallel utterances for testing, but the contexts were different from those of the HUB corpus. Therefore, the total number of source-target pairs in the SPOKE task was 16 (four SPOKE source speakers × four HUB target speakers), which included four female-to-female (F-F) pairs, four female-to-male (F-M) pairs, four male-to-female (M-F) pairs, and four male-to-male (M-M) pairs. Furthermore, the VCC2018 organizer also provided the transcripts of the corpus. The sampling rate of speech signals was set to 22050 Hz and the resolution per sample was 16 bits.

The reference speech used to construct the cascade VC system was generated by a concatenative unit-selection TTS system, which was trained by around 3000 utterances from a single male speaker. Notably, although the linguistic contexts were the same, each speaker still had different prosody, such as short-pause positions, which resulted in significantly different spectral characteristics. To alleviate these acoustic mismatches, we controlled the short-pause positions of the TTS voices so that they corresponded to those of the individual source and target speakers.

Table 1. *MCD scores of source and DNN VC systems* ***w/o GV*** *(F: female, M: male).*

	Source	*OtoO*	*wRTTS*	*wRspk*
F - F	8.27	**5.37**	5.54	5.73
F - M	8.46	**5.51**	5.66	5.67
M – F	8.46	**5.54**	5.68	5.67
M –M	7.89	**5.44**	5.65	5.63
Avg.	8.33	**5.48**	5.64	5.67

Table 2. *MCD scores of source and DNN VC systems* ***w/ GV*** *(F: female, M: male).*

	Source	*OtoO*	*wRTTS*	*wRspk*
F - F	8.27	**5.94**	6.09	6.11
F - M	8.46	**6.18**	6.29	6.24
M – F	8.46	**6.16**	6.23	6.18
M –M	7.89	**6.11**	6.23	6.21
Avg.	8.33	**6.12**	6.22	6.19

The multi-speaker WaveNet vocoder was trained by the data from all speakers of the VCC2018 corpus and speakers “bdl” and “slt” of the CMU-ARCTIC [31] corpus. The number of training utterances was 81 per speaker in the VCC2018 corpus and 1132 per speaker in the CMU-ARCTIC corpus. The total data length was about three hours. Moreover, the speaker-dependent WaveNet vocoders were constructed by using the training data of each target speaker to update the output layers of the multi-speaker WaveNet vocoder.

The feature extraction and analysis-synthesis framework were based on the WORLD vocoder. We used WORLD to extract a 513-dimensional spectral envelope, 513-dimensional *ap*, and one-dimensional F_0 with 25 ms frame length and 5 ms frame shift. The spectral envelope was parameterized into a 34-dimensional mel-cepstrum, and *ap* was coded into a two-dimensional aperiodic component. Joint spectral features were constructed by dynamic time warping (DTW) based on the corresponding mel-cepstrum. The mel log spectrum approximation (MLSA) filter [32] was used as the synthesis filter of the DIFFVC process in the analysis-synthesis framework.

Both GMM-based and DNN-based VC were conducted in the internal evaluations. The hyperparameters of our feed-forward neural network were set as follows: four hidden layers with 1024 hidden units per layer. The nonlinear activation function was rectified linear unit (ReLU) and the optimization algorithm was Adam [33]. The weights were randomly initialized by Xavier [34] and the biases were initially set to zero. The learning rate was 0.0006, the number of training epochs was 15, and the utterance mini-batch was used. On the other hand, the settings of the baseline GMM were 32 mixtures and a full covariance matrix.

Our WaveNet vocoder consisted of 30 connected residual blocks, which included one layer comprising of a dilated causal convolution, a gate activated function, and one residual per residual block. The total number of dilated causal convolution channels was 512, and the dilations of 30 layers were set to 3 sets of [2^0, 2^1, 2^2, ..., 2^9]. The 1×1 convolutions in the residual

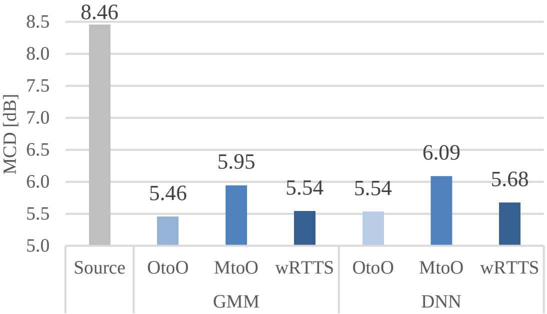

Figure 5. *MCD scores of male-to-female pairs based on GMM VC and DNN VC systems **w/o GV**.*

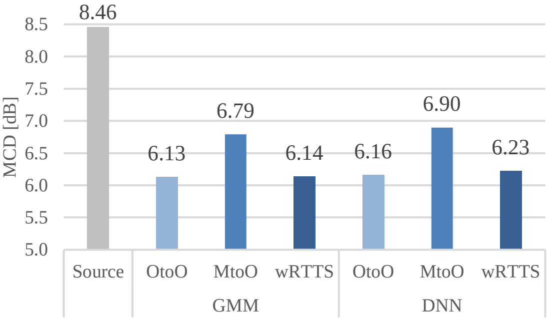

Figure 6. *MCD scores of male-to-female pairs based on GMM VC and DNN VC systems **w/ GV**.*

block were set to 512 channels, and the number of 1×1 convolution channels between the skip-connection and the softmax layer was 256. Adam was also used for optimization, and its learning rate was initialized as 0.001 with 50% decay per 50,000 iterations. The mini-batch size was set to 20,000 samples and the number of iterations was 200,000.

4.2. Reference speaker evaluation

According to the results of [11], the performance when choosing parallel data for non-parallel training is almost the same as that for non-parallel data, and it is easy to compare a non-parallel training system with a parallel training system using only a parallel corpus. Therefore, we evaluated the proposed cascade VC system on the basis of the 12 speaker pairs formed by the four SPOKE speakers. The objective measurement was the mel-cepstrum distance (MCD), which is defined as

$$MelCD(\text{conv},\text{tar})[\text{dB}] = \frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{I}\left(m_i^{(\text{conv})} - m_i^{(\text{tar})}\right)^2}, \quad (7)$$

where $m_i^{(\text{conv})}$ are the mel-cepstral coefficients of converted features, $m_i^{(\text{tar})}$ are the mel-cepstral coefficients of actual target features, and i is the dimension of the mel-cepstral coefficients.

To verify the effectiveness of the proposed cascade VC system and the effect of the unnaturalness from TTS-generated speech, we first compared the following three DNN based VC systems:

- **OtoO**: the basic **one-to-one** parallel VC system.
- **wRTTS**: the proposed cascade VC system **with the TTS output as the reference speech**.

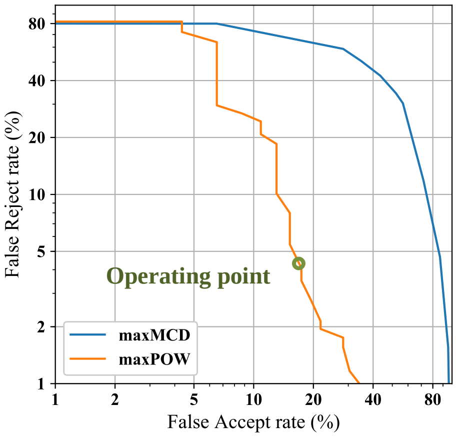

Figure 7. *DET curves of the two detectors.*

- **wRspk**: the proposed cascade VC system **with natural speech as the reference speech**.

Specifically, we used the SPOKE speaker rather than the source and target speakers as the reference speaker in **wRspk**. From Tables 1 and 2, the MCDs of **wRTTS** are almost the same as those of **wRspk**; thus, the results show that TTS output already contains sufficient acoustic components to be used as the reference speech. Furthermore, the MCD differences between **wRTTS** and **OtoO** are about 0.2 dB, indicating that the mismatch between the training and converted reference features still causes performance degradation.

Next, we compared the proposed cascade VC framework with the baseline non-parallel VC system (**MtoO**) provided by the VCC2018 organizer [35], which used two speakers and a target speaker to train a speaker-independent (**many-to-one**) VC model. We constructed a gender-dependent, male-to-female, and speaker-independent VC model by using one male speaker and TTS output. The evaluation involved cross-validation of the two male speakers in the SPOKE set. Moreover, the baseline system was GMM-based VC; thus, we evaluated the performance of both GMM-based and DNN-based models. As shown in Figs. 5 and 6, the proposed system outperforms the **MtoO** systems for both the GMM-based and DNN-based models. To summarize, for the speaker spectral conversion, the proposed method achieves a comparable performance to the **OtoO** system and outperforms the baseline **MtoO** system in the objective evaluation. In addition, because the potential ability for joint optimization with the WaveNet vocoder, we adopted the DNN-based spectral conversion in our final submitted system.

4.3. Evaluation of system selection

The goal of system selection is to detect utterances with collapsed speech segments, and then to replace them by utterances without collapsed speech. Therefore, we evaluated the system selection by measuring how many collapsed utterances had been detected. That is, we assumed collapsed speech detection as a verification problem; thus, the performance of the detector was measured by the false accept and false reject rates.

For the evaluation dataset, a human subject labeled the converted utterances of all source-target pairs in the SPOKE task, which were generated by the WaveNet vocoder on the

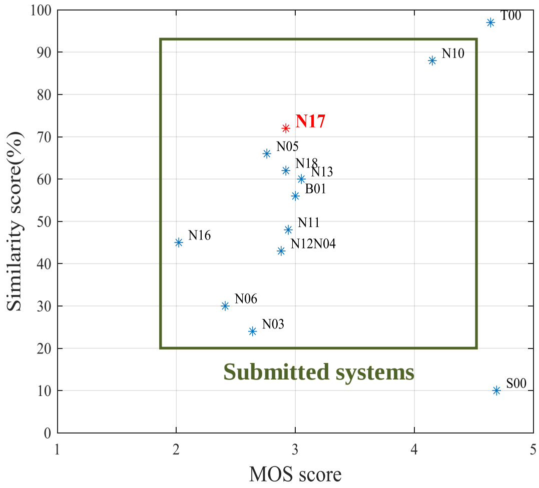

Figure 8. *Overall summary of evaluation results for VCC2018 SPOKE task*

basis of the auxiliary features of the converted F_0, coded *ap*, and converted spectral feature without the analysis-synthesis framework. The number of converted utterances was 560 and the number of the labeled collapsed utterances was 47. Two detectors were compared:

- **maxMCD**: A voice activity detection system (VAD) was first applied to all generated utterances, and then the MCDs of WaveNet-generated utterances and WORLD-generated utterances where calculated. The final score was the maximum difference between the WaveNet MCDs and WORLD MCDs.
- **maxPOW**: The proposed measurement is the difference between the maximum power of the WaveNet-generated utterance and the WORLD-generated utterance.

Figure 7 shows the detection error tradeoff (DET) curves of the two detectors. Not only the equal error rate (EER) of **maxPOW** is much lower than that of **maxMCD**, but also the entire curve for **maxPOW** is lower than that for **maxMCD**. The results indicate that the proposed **maxPOW** score is a more robust measurement of collapsed speech detection than the **maxMCD** score, and we can detect 80% of collapsed utterances with a false reject rate of less than 5% for clean utterances.

4.4. External evaluation results from VCC2018

The VCC2018 organizer conducted subjective tests on all submitted systems for both the HUB task and the SPOKE task. The evaluations included naturalness and similarity tests. In the naturalness tests, the measurement was the five-point mean opinion score (MOS), where “5” stood for “completely natural” and “1” stood for “completely unnatural”. In the similarity tests, listeners were asked to decide whether or not the converted utterances and target utterances were spoken by the same person. A four point scale was given to listeners: “definitely the same”, “probably the same”, “probably different” and “definitely different”. The final similarity scores were the percentage of the summation of “definitely the same” and “probably the same”.

Table 3. *p-values of the naturalness evaluation results of VCC2018 SPOKE task.*

null hypothesis	Quality is better or worse than N17
p-value > 0.6	Baseline, N11, N18, N04, N12
p-value = 0.073	N13
p-value = 0.016	N05
p-value < 2^{-16}	Source, Target, N10, N06, N16

For the submitted NU non-parallel VC system, each decoder was trained using all the training data of the corresponding SPOKE source speaker and TTS outputs, and each Decoder was trained using all training data of the corresponding HUB target speaker and TTS outputs. Four modes of generated speech were used as the candidates in the system selection:

- **WN-diff-anasyn**: speech generated by the WaveNet vocoder on the basis of the converted F_0, coded *ap*, and converted spectral features with analysis-synthesis framework (f_0‘, coded a_2, and m_3 in Fig.1).
- **WN-diff-anasyn-lpc**: speech generated under the same conditions as **WN-diff-anasyn**, but the conditional probabilities of the WaveNet vocoder were constraint on previous samples by linear predictive coding.
- **WN-diff**: speech generated under the same conditions as **WN-diff-anasyn**, but *ap* and the converted spectral features were processed without the last synthesis-analysis step (a_1 and m_2 in Fig. 1).
- **WD-diff-anasyn**: speech generated under the same conditions as **WN-diff-anasyn**, but with the WaveNet vocoder replaced by the WORLD vocoder.

The priority was set as 1. **WN-diff-anasyn**, 2. **WN-diff-anasyn-lpc**, 3. **WN-diff**, and 4. **WD-diff-anasyn** according to an unofficial internal evaluation. Specifically, we performed collapsed speech detection of all utterances generated by the WaveNet vocoder, and then the system selected the final submitted files according to the detection results and the predefined priority. Furthermore, the detection threshold was set on the basis of the operating point from the DET curve in Fig. 7 corresponding to a 5% false reject rate and 20% false accept rate. The final submitted files contained 3% **WN-diff-anasyn-lpc**, 1% **WN-diff**, and 1% **WD-diff-anasyn** files.

Figure 8 shows the overall results and Table 3 demonstrates the significance relationships of our system (N17) with others in terms of the p-values in the naturalness evaluations. Our system is about third place in the naturalness evaluations and second place in the similarity measurements. The average MOS of the proposed system is about 3, and the average similarity accuracy is about 70%, as described in detail in the following.

4.4.1. Naturalness

As shown in Table 4, the MOS scores of the proposed VC system are stable for each pair, indicating the effectiveness of the proposed system under different conversion conditions. Compared with the baseline system (B01), the results of cross-gender evaluations are consistent with the objective evaluations (Figs. 5 and 6), implying that the spectrum prediction of our VC system is better than that of the speaker independent GMM VC

Table 4. *Naturalness evaluation results of VCC2018 SPOKE task. (F: female, M: male).*

System	F-F	F-M	M-F	M-M	Avg.
Source	**4.69**	**4.69**	**4.69**	**4.69**	**4.69**
Target	**4.64**	**4.64**	**4.64**	**4.64**	**4.64**
N10	**4.24**	**4.19**	**4.02**	**4.15**	**4.15**
N13	3.12	**3.05**	**2.90**	**3.11**	**3.05**
Baseline	**3.60**	2.66	2.46	**3.29**	**3.00**
N11	**3.28**	2.83	**2.93**	2.73	**2.94**
N18	**3.25**	2.77	**2.94**	2.72	**2.92**
N17	**3.20**	**2.86**	**2.75**	**2.85**	**2.92**
N04	2.89	**2.93**	2.69	**3.03**	2.88
N12	**3.38**	**3.05**	2.08	**3.00**	2.88
N05	**3.20**	2.49	2.56	2.82	2.76
N03	2.70	2.81	2.13	**2.92**	2.64
N06	2.93	2.21	2.05	2.46	2.41
N16	2.20	1.93	1.82	2.13	2.02

Table 5. *Similarity evaluation results of VCC2018 SPOKE task. (F: female, M: male).*

System	F-F	F-M	M-F	M-M	Avg.
Source	10%	10%	10%	10%	10%
Target	**97%**	**97%**	**97%**	**97%**	**97%**
N10	**83%**	**94%**	**74%**	**86%**	**88%**
N17	**79%**	**71%**	**70%**	**59%**	**72%**
N05	65%	68%	56%	**78%**	66%
N18	66%	**72%**	37%	55%	62%
N13	55%	66%	53%	**70%**	60%
Baseline	66%	59%	49%	35%	56%
N11	46%	60%	18%	50%	48%
N16	38%	63%	14%	56%	45%
N12	25%	60%	9%	**68%**	43%
N04	24%	69%	17%	57%	43%
N06	28%	33%	0%	50%	30%
N03	17%	37%	3%	34%	24%

system. However, the proposed VC system exhibited worse performance in the intra-gender task, because the baseline system used the vocoder-free framework in the intra-gender conversion instead of the conventional vocoder framework in the inter-gender conversion [35]. To be more specific, the results indicate that the WaveNet vocoder still suffers from broken excitation signals. Because excitation signals extracted by the conventional vocoder usually suffer from serious distortion, although we used the WaveNet vocoder for synthesis instead of WORLD, the WaveNet vocoder still suffers from broken excitation signals extracted from WORLD. In addition, we find that our collapsed speech detection technique can only detect extremely white noise, whereas the WavNet vocoder sometimes generates short impulse noises, and which have a significant effect on human auditory perception. Therefore, we may improve the quality of converted speech by solving the problem of the broken excitation signal so that the WaveNet vocoder generates a stable output.

4.4.2. Similarity

From the results in Table 5, we find that our VC system outperforms the baseline system for both cross-gender and same-gender tasks, showing that the WaveNet vocoder can retain the characteristic of the target's timbre better than both the WORLD and vocoder free DIFFVC frameworks. Nonetheless, although our system achieves an above-average accuracy for similarity in cross-gender and F-F pairs, the performances of our M-M pairs are seriously degraded. Broken excitation signals may also cause performance degradation because WORLD often has the difficulty in extracting the correct F_0 for male speakers. We also find that the WaveNet vocoder is more sensitive to incorrectly predicted F_0 than WORLD, resulting in our M-M and M-F conversion sets containing many utterances with scratchy sounds. The scratchy sounds usually cause significant blurring of speaker identity, particularly in same-gender cases; thus, the M-M set has significantly more degradation than other sets of the proposed system. Moreover, although the M-F set of our system exhibits less degradation than the M-M set in similarity tests, it still achieves the worst naturalness performance among the conversion sets of our system as shown in Table 4. As a result, compensation of the broken extracted F_0 or bypassing conventional vocoder analysis may be the key to improve both the quality and similarity of converted speech.

5. CONCLUSION

In this paper, we describe the details of the NU non-parallel VC system developed for VCC2018. The main concept is the use of TTS outputs as a bridge to connect non-parallel source and target speaker utterances. Furthermore, we also propose a system selection technique to automatically select the collapsed-speech-free utterances from different generated conditions and vocoders. Internal experimental results reveal that the proposed VC system can achieve comparable spectrum prediction accuracy to a parallel VC system as well as the effectiveness of the system selection technique. In addition, the subjective evaluations provided by the VCC2018 organizer demonstrate that our VC system achieves an above average performance in both quality and similarity measurements. As future work, to further improve the quality of converted speech, we intend to study the techniques of robust collapsed speech detection and the compensation of unnatural excitation signals.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

1. A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proc. ICASSP*, vol. 1, pp. 285-288, 1998.

2. T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. ASLP*, vol. 15, no. 8, pp. 2222-2235, Nov. 2007.

3. D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Trans. ASLP*, vol. 18, no. 5, pp. 922-931, July 2010.

4. E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling,

for parallel or nonparallel corpora," *IEEE Trans. ASLP*, vol. 20, no. 4, pp. 1313-1323, May 2012.

5. S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Trans. ASLP*, vol. 18, no. 5, pp. 954-964, July 2010.
6. 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. ASLP*, vol. 22, no. 12, pp. 1859-1872, Dec. 2014.
7. T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using deep belief nets," *Proc. INTERSPEECH*, pp. 369-372, 2013.
8. R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," *Proc. Spoken Language Technology Workshop (SLT)*, 2012.
9. Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Trans. ASLP*, vol. 22, no. 10, pp. 1506-1521, 2014.
10. Y.-C. Wu, H.-T. Hwang, C.-C. Hsu, Y. Tsao, and H.-M. Wang, "Locally linear embedding for exemplar-based spectral conversion," *Proc. INTERSPEECH*, pp. 1652-1656, 2016.
11. D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Trans. ASLP*, vol. 18, no. 5, pp. 944-953, July 2010.
12. 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," *Proc. APSIPA*, pp. 1-6, 2016.
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 adersarial networks," *Proc. INTERSPEECH*, 2017.
14. L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," *Proc. ICME*, pp. 1-6, 2016.
15. F. Xie, F. K. Soong, and H. Li, "A KL divergence and DNN-based approach to voice conversion without parallel training sentences," *Proc. INTERSPEECH*, pp. 287–291, 2016.
16. T. Toda, Y. Ohtani, and K. Shikano "One-to-many and many-to-one voice conversion based on eigenvoices," *Proc. ICASSP*, vol. 4, pp. 1249-1252, 2007.
17. Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Many-to-many eigenvoice conversion with reference voice," *Proc. INTERSPEEH*, pp. 1623-1626, 2009.
18. T. Masuda and M. Shozakai, "Cost reduction of training mapping function based on multistep voice conversion," *Proc. ICASSP*, vol. 3, pp. 693–696, 2007.
19. H. Doi, T. Toda, T. Nakano, M. Goto, and S. Nakamura, "Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system," *Proc. APSIPA*, 2012.
20. 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.
21. A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," *Proc. INTERSPEECH*, pp. 1118–1122, 2017.
22. K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with WaveNet-based waveform generation," *Proc. INTERSPEECH*, pp. 1138–1142, 2017.
23. T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," *Proc. ASRU*, 2017.
24. K. Kobayashi, S. Takamichi, S. Nakamura, and T. Toda, "The NU-NAIST voice conversion system for the Voice Conversion Challenge 2016," *Proc. INTERSPEECH*, pp. 1667-1671, 2016.
25. A. van den 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," *CoRR abs/1609.03499*, 2016.
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. M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Commun.*, vol. 84, pp. 57-65, Nov. 2016.
28. H. T. Hwang, Y. Tsao, H. M. Wang, Y. R. Wang and S. H. Chen, "A probabilistic interpretation for artificial neural network-based voice conversion," *Proc. APSIPA*, 2015.
29. P. Lumban Tobing, H. Kameoka, and T. Toda, "Deep acoustic-to-articulatory inversion mapping with latent trajectory modeling," *Proc. APSIPA*, 2016.
30. K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," *Proc. ICASSP*, pp. 660–663, 1995.
31. J. Kominek and A. W. Black, "The CMU ARCTIC speech databases for speech synthesis research," *Tech. Rep. CMU-LTI-03-177*, 2003.
32. K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," *Proc. ICSLP*, pp. 1043-1045, 1994.
33. D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proc. ICLR*, 2015.
34. X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Proc. AISTATS*, vol. 9, pp. 249-256, 2010.
35. K. Kobayashi and T. Toda, "Sprocket: open-source voice conversion software," *Submitted to Odyssey*, 2018.