Hybrid Text-to-Speech for Articulation Disorders with a Small Amount of Non-Parallel Data

Ryuka Nanzaka and Tetsuya Takiguchi
Kobe University, Japan
E-mail: nanzaka@me.cs.scitec.kobe-u.ac.jp, takigu@kobe-u.ac.jp

***Abstract*—Conventional approaches to statistical parametric speech synthesis usually require a large amount of speech data. But it is very difficult for persons with articulation disorders, in particular, to utter a large amount of speech data, and their utterances are often unstable or unclear so that we cannot understand what they say. In this paper, we propose a hybrid approach for a person with an articulation disorder, using two models of a physically unimpaired person and a person with an articulation disorder to generate an intelligible voice while preserving the speaker's individuality (with an articulation disorder). Our method has two processes - the speech synthesis part and voice conversion part. Speech synthesis is employed for obtaining a speech signal (of a physically unimpaired person), where a large amount of training data of a physically unimpaired person is used. Then, voice conversion (VC) is employed for converting the voice of the physically unimpaired person to that of a person with an articulation disorder, where a small amount of speech data of a person with an articulation disorder is only used for training VC. Also, a cycle-consistent adversarial network (CycleGAN) that does not require parallel data is employed for VC. An objective evaluation showed that the mel-cepstrum obtained using our method are close to the target in terms of global variance (GV) and modulation spectrum (MS).**

I. Introduction

In this study, we focus on persons with articulation disorders resulting from the athetoid type of cerebral palsy. In the case of persons with articulation disorders resulting from the athetoid type of cerebral palsy, his/her movements are sometimes more unstable than usual. That means their utterances (especially their consonants) are often unstable or unclear due to their athetoid symptoms, and there is a great need for voice systems that can assist them in their communication [1].

Text-to-speech synthesis is a technique of synthesizing speech from arbitrarily given text. Many methods for realizing text-to-speech synthesis have been proposed so far. The most representative method was an approach based on the hidden Markov model (HMM) [2]. Recently, an approach based on deep neural networks (DNN) [3] has come into mainstream use for speech synthesis because it generates better sound than the conventional hidden Markov model approach. Also, speech synthesis is used to support persons with disabilities. For example, Yamagishi *et al.* [4] collected voice of various persons to construct a text-to-speech (TTS) system for amyotrophic lateral sclerosis (ALS) patients.

Voice conversion [5] is a technique to convert information such as speaker character and emotion, while maintaining the contents of an utterance with respect to input speech. A Gaussian mixture model (GMM) [6] is a typical technique for statistical voice conversion. In recent years, there are also many models that make use of DNN due to the growing interest in deep learning. Parallel data is required for these voice conversion methods, but there are some approaches that do not use parallel data, such as Boltzmann machine (adaptive restricted Boltzmann machine) [7], CycleGAN [8], and others. These models solve the quality problem caused by the mismatching of time-alignment for parallel data.

As mentioned previously, it is very difficult for a person with an articulation disorder to utter a large amount of speech data due to the severe physical strain of cerebral palsy. In fact, it may be difficult to even recognize what they say. Also, the mismatching of the time-alignment of the parallel utterances between persons with articulation disorders and those who are physically unimpaired is not negligible because it may not be so easy for persons with articulation disorders to utter some phonemes (phoneme deletion problem). In this paper, we propose a TTS system for a person with an articulation disorder using a small amount of speech data with the goal of making clear voices while preserving the speaker's individuality (with an articulation disorder).

A simple TTS system for a person with an articulation disorder may be trained using only his/her utterances. But the synthesized speech is unstable or unclear so it is difficult to understand what they are saying. Also, it requires a large amount of training data, but it is very difficult for a person with an articulation disorder to utter a large amount of training data. Therefore, in this paper, a hybrid TTS system of a physically unimpaired person and a person with an articulation disorder (for voice conversion) is proposed. A TTS system for a physically unimpaired person is employed for synthesizing a speech signal and then the synthesized speech signal is converted to that of a person with an articulation disorder.

II. Hybrid TTS system

The flow of our proposed method is shown in Fig. 1. Our proposed approach is a hybrid TTS between TTS and VC. First a TTS system using the bidirectional long short-term memory (LSTM) that was trained using speech data of a physically unimpaired person is employed for synthesizing a speech signal. Then the synthesized speech is converted to that of a person with an articulation disorder using VC based on CycleGAN.

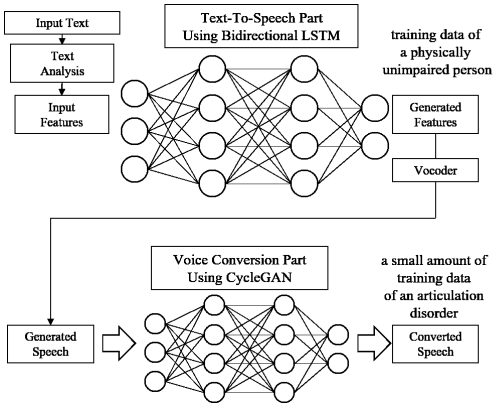


Fig. 1. A flow of our proposed method: hybrid of TTS and VC

A. TTS using Bidirectional LSTM

The flow of TTS using Bidirectional LSTM [9] is shown in Fig. 2. Bidirectional LSTM is based on recurrent neural networks (RNN). It computes a hidden state vector sequence $\boldsymbol{h} = (h_1, \cdots, h_T)$ and outputs a vector sequence $\boldsymbol{y} = (y_1, \cdots, y_T)$, for a given input vector sequence $\boldsymbol{x} = (x_1, \cdots, x_T)$, iterating the following equations from $t = 1$ to T:

$$h_t = H(\boldsymbol{W}_{xh} x_t + \boldsymbol{W}_{hh} h_{t-1} + b_h) \quad (1)$$

$$y_t = \boldsymbol{W}_{hy} h_t + b_y \quad (2)$$

where $\mathbf{W}$ is the weight matrices (e.g. $\mathbf{W_{xh}}$ is the weight matrix between input and hidden vectors); b is the bias vectors (e.g. b_h is the bias vector for hidden state vectors); and H is the nonlinear activation function for hidden nodes. Usually, a sigmoid or hyperbolic tangent function is used, but it sometimes causes a gradient vanishing problem. Therefore, in LSTM, H is implemented with the following functions:

$$i_t = \sigma(\boldsymbol{W}_{xi} x_t + \boldsymbol{W}_{hi} h_{t-1} + \boldsymbol{W}_{ci} c_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma(\boldsymbol{W}_{xf} x_t + \boldsymbol{W}_{hf} h_{t-1} + \boldsymbol{W}_{cf} c_{t-1} + b_f) \quad (4)$$

$$c_t = f_t c_{t-1} + i_t tanh(\boldsymbol{W}_{xc} x_t + \boldsymbol{W}_{hc} h_{t-1} + b_c) \quad (5)$$

$$o_t = \sigma(\boldsymbol{W}_{xo} x_t + \boldsymbol{W}_{ho} h_{t-1} + \boldsymbol{W}_{co} c_t + b_o) \quad (6)$$

$$h_t = o_t tanh(c_t) \quad (7)$$

where σ is the sigmoid function; i, f, o, and c are input gate, forget gate, output gate and cell memory, respectively. Bidirectional RNN has a forward state sequence, $\overrightarrow{h}$, and a backward state sequence, $\overleftarrow{h}$ as follows:

$$\overrightarrow{h_t} = H(\boldsymbol{W}_{x\overrightarrow{h}} x_t + \boldsymbol{W}_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}) \quad (8)$$

$$\overleftarrow{h_t} = H(\boldsymbol{W}_{x\overleftarrow{h}} x_t + \boldsymbol{W}_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}) \quad (9)$$

$$y_t = \boldsymbol{W}_{\overrightarrow{h}y} \overrightarrow{h}_t + \boldsymbol{W}_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y \quad (10)$$

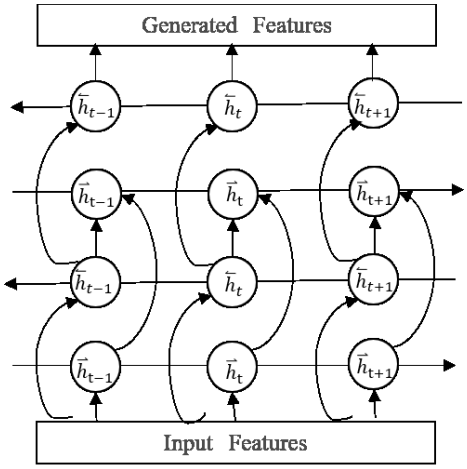


Fig. 2. Bidirectional LSTM speech synthesis system

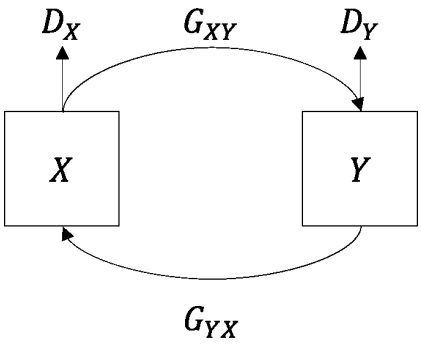


Fig. 3. A flow of CycleGAN

B. Voice Conversion using CycleGAN

CycleGAN consists of two generators (G_{XY}, G_{YX}) and two discriminators (D_X, D_Y) as shown in Fig. 3. They are trained with adversarial loss, cycle-consistency loss, and identity-mapping loss [10].

1) Adversarial Loss: Adversarial loss is shown as follows:

$$L_{adv}(G_{XY}, D_Y) = E_{y \sim P_{Data(y)}}[logD_Y(y)] + E_{x \sim P_{Data(x)}}[log(1 - D_Y(G_{XY}(x)))] \quad (11)$$

Generator, G_{XY} generates data that cannot distinguish from the target y by minimizing this function. On the other hand, D_Y maximizes this function to distinguish the generated data from target y. Therefore, the distribution of the generated data approaches the distribution of the target data. Inverse transform, G_{YX} is the same. It is known that adversarial loss is effective at avoiding statistical averaging. This makes it possible to alleviate the problem of over-smoothing, which is one of the major causes of quality degradation of speech obtained by a statistical method.

2) Cycle-Consistency Loss: The cycle-consistency loss function is analogous to the objective function of an autoencoder, which minimizes the difference between the input and output to reconstruct the input from the output. Thus, the cycle-consistent loss is defined as follows:

$$L_{cyc}(G_{XY}, G_{YX}) = E_{y\sim P_{Data(x)}}[||G_{YX}(G_{XY}(x)) - x||_1] + E_{x\sim P_{Data(y)}}[||G_{XY}(G_{YX}(y)) - y||_1] \quad (12)$$

This loss function gives a structural restriction of returning to the original by inversion after conversion. It enables the model to keep the context of the input.

3) Identity-Mapping Loss: A cycle-consistency loss is not enough to guarantee that the mappings always preserve linguistic information. To maintain linguistic-information preservation without relying on extra modules, an identity-mapping loss is defined as follows:

$$L_{id}(G_{XY}, G_{YX}) = E_{y\sim P_{Data(y)}}[||G_{XY}(y) - y||_1] + E_{x\sim P_{Data(y)}}[||G_{YX}(x) - x||_1] \quad (13)$$

When y is the input, the generator G_{XY} does nothing. G_{YX} is the same. When this loss function is fulfilled, the generator works well as the function to convert the input feature into the target one.

By combining these losses, a model can be learned from unpaired training samples, and the learned mappings are able to map an input x (or y) to a desired output y (or x). The combined loss is shown as follows:

$$\begin{aligned} L_{all}(G_{XY}, G_{YX}, D_X, D_Y) &= L_{adv}(G_{XY}, D_Y) + L_{adv}(G_{YX}, D_X) \\ &\quad + \lambda_1 L_{cyc}(G_{XY}, G_{YX}) + \lambda_2 L_{id}(G_{XY}, G_{YX}) \end{aligned} \quad (14)$$

where λ controls the relative importance of the three losses. In this work, one-dimensional convolutional neural networks (CNN) are used to model the generators and discriminators.

III. Experimental Setup

The utterances of a Japanese male with an articulation disorder and a physically unimpaired male person are used. The scripts for utterances are based on the ATR Japanese speech database [11]. 450 training utterances are used in Bidirectional LSTM and 100 training utterances are used for CycleGAN.

The audio data were sampled at 16 kHz with a bit depth of 16 bits and shifted every 5ms. The 60 mel-frequency cepstral coefficients, f_0, and aperiodicity were extracted by using STRAIGHT [12]. When we trained the CycleGAN, we cropped a fixed-length segment (128 frames) randomly from a randomly selected audio file.

The input features for the LSTM are 975 linguistic contexts, and the input features for the CycleGAN are 60 mel-frequency cepstral coefficients. The aperiodicity is kept unmodified. When carrying out conversion, f_0 are modified by using the linear mean-variance transformations as follows:

$$f'_{0_t} = \frac{\sigma_{trg}}{\sigma_{src}}(f_{0_t} - \mu_{src}) + \mu_{trg} \quad (15)$$

f'_{0_t} is the f_0 after conversion. μ_{trg} and σ_{trg} are the mean and variance of the f_0 of an utterance of a person with an articulation disorder, respectively. μ_{src} and σ_{src} are the mean and variance of the f_0 of a physically unimpaired person, respectively. f_{0_t} is the f_0 of an articulation disorder at frame t.

As a baseline method, we use the speech data that were directly synthesized using Bidirectional LSTM with the speech data of a person with an articulation disorder, which is called as DNN in Figs. 5, 6, and 7.

IV. Experimental Results

A. Comparison of Spectrograms

The spectrograms are shown in Fig. 4. The spectrogram of articulation disorders has a tendency to lose high-frequency power as shown in (b). Also, the generated spectrogram has a similar tendency, especially over 2,000 Hz. Therefore, the hybrid model successfully captures the spectrum features of articulation disorders.

In our method, to improve the unstable f_0, the estimated f_0 patterns of a physically unimpaired person are used for the unstable pitch (of the articulation disorder), with the average f_0 being converted to the target f_0 in VC (Eq. (15)). Also, the duration of a person with an articulation disorder in (b) is slower than that of the physically unimpaired person in (a), which cause their speech to be less intelligible [13], [14]. In our method, the duration patterns of the physically unimpaired person are used so that the intelligibility of the generated sound improves.

B. Global Variance

Global variance (GV) [15] is widely known as a measure that quantitatively explains over-smoothing. Fig. 5 shows each global variance. As shown in this figure, the GV of our method is similar to that of the target without using parallel data.

C. Modulation Spectrum

Fig. 6 shows each modulation spectrum [16]. As shown in this figure, there is little difference between the baseline and proposed method, and they are close to the target spectrum.

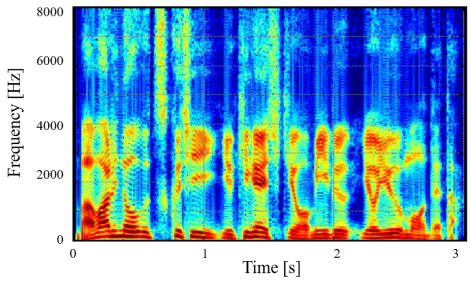
D. Mel-cepstral Distortion

Mel-cepstral distortion is calculated as follows:

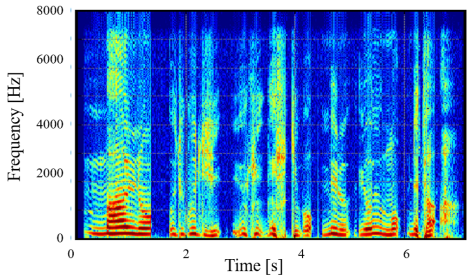
$$\mathrm{MelCD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{P}(mc_i^{(x)} - mc_i^{(y)})^2} \quad (16)$$

where $mc_i^{(x)}$ is the mel-cepstral coefficients of the target speech, and $mc_i^{(y)}$ is the mel-cepstral coefficients of the converted speech.

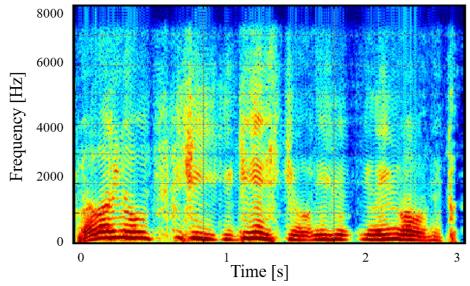
The comparison of the mel-cepstral distortion is shown in Fig. 7. As shown in this figure, our method outperformed the baseline. The mean distortion of our method and baseline was



(a) Source spectrogram (a physically unimpaired person)



(b) Target spectrogram (a person with an articulation disorder)



(c) Generated spectrogram

Fig. 4. Comparison of spectrograms

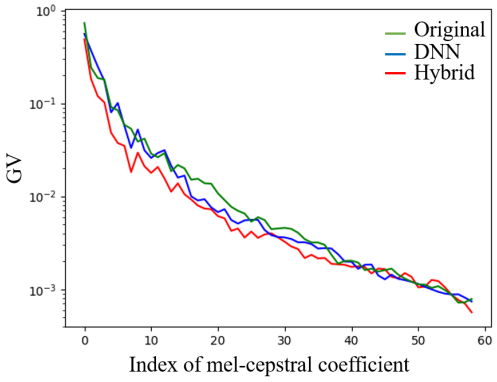


Fig. 5. Comparison of global variance

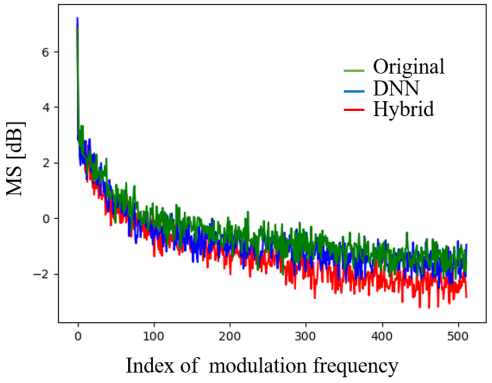


Fig. 6. Comparison of modulation spectrum

7.66 and 10.65, respectively. The baseline system was trained using 100 sentences, but it was not enough to estimate the features, such as the duration. Our method, however, deals with this problem completely by using two models.

In our method, the generated sound consists of the correct phonemes sequence without deletion because the (synthesized) spectrum of a physically unimpaired person is converted to that of a person with an articulation disorder using VC. On the other hand, as a person with an articulation disorder may not be able to utter some phonemes, it is difficult to train the baseline system correctly using only speech data of a person with an articulation disorder.

V. Conclusions

In this paper, we proposed a speech synthesis method using a small amount of training data from a person with an articu-

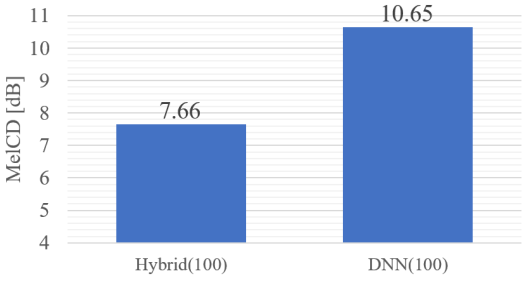


Fig. 7. Comparison of MelCD

lation disorder by combining TTS using Bidirectional LSTM and voice conversion using CycleGAN. The comparison of spectrograms showed that the high-frequency features and the formant of the low-frequency in the spectrogram of articulation disorders are well learned.

An objective evaluation showed that the mel-cepstrum sequences obtained with our method are close to the target in terms of global variance and modulation spectrum.

In the future, we will decide the optimum weight of the loss function and consider introducing another loss function in CycleGAN. Then, we intend to investigate speaker character and clarity by subjective evaluation. Also, because the degradation of speech is inevitable when it is passed through a vocoder, we would also like to consider end-to-end, vocoder-free synthesis methods.

REFERENCES

[1] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "A preliminary demonstration of exemplar-based voice conversion for articulation disorders using an individuality-preserving dictionary," in *EURASIP Journal on Audio, Speech, and Music Processing*, 2014.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *ICASSP*, pp. 1315-1318, 2000.

[3] H. Zen, A. Senior, and M. Schuster, "Statistical Parametric Speech Synthesis Using Deep Neural Networks," in *ICASSP*, pp. 7962-7966, 2013.

[4] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities:Voice banking and reconstruction," in *Acoustical Science and Technology*, vol. 33, no. 1, pp. 1-5, 2012.

[5] S. Mohammadi and A. Kain, "An overview of voice conversion systems," in *Speech Communication*, vol. 88, no. 1, pp. 65-82, 2017.

[6] T. Toda, A. Black, and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222-2235, 2007.

[7] T. Nakashika, T. Takiguchi, and Y. Minami, "Non-Parallel Training in Voice Conversion Using an Adaptive Restricted Boltzmann Machine," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2032-2045, 2016.

[8] T. Kaneko and H. Kameoka, "Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks," *arXiv*, 2017.

[9] Y. Fan, Y. Qian, F. Xie, and Frank K. Soong, "TTS Synthesis with Bidirectional LSTM based Recurrent Neural Networks," *INTERSPEECH*, pp. 1964-1968, 2014.

[10] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *ICLR*, 2017.

[11] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T. Umeda, and H. Kuwahara, "A large-scale Japanese speech database," in *ICSLP*, pp. 1089-1092, 1990.

[12] H. Kawahara, I. Masuda, and A. 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," in *Speech Communication*, vol. 27, no. 3-4, pp. 187-207, 1999.

[13] R. Ueda, T. Takiguchi, and Y. Ariki, "Individuality-Preserving Voice Reconstruction for Articulation Disorders Using Text-to-Speech Synthesis," in *ACM ICMI*, pp. 343-346, 2015.

[14] R. Aihara, T. Takiguchi, and Y. Ariki, "Phoneme-Discriminative Features for Dysarthric Speech Conversion," in *Interspeech*, pp. 3374-3378, 2017.

[15] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," in *IEICE Trans.*, vol. E90-D, no. 5, pp. 816-824, 2007.

[16] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify the modulation spectrum in HMM-based speech synthesis," in *ICASSP*, pp. 290-294, 2014.