# HIGH-INTELLIGIBILITY SPEECH SYNTHESIS FOR DYSARTHRIC SPEAKERS WITH LPCNET-BASED TTS AND CYCLEVAE-BASED VC

*Keisuke Matsubara*[1,2*], *Takuma Okamoto*[2], *Ryoichi Takashima*[1], *Tetsuya Takiguchi*[1], *Tomoki Toda*[3,2], *Yoshinori Shiga*[2], *and Hisashi Kawai*[2]

[1]Graduate School of System Informatics, Kobe University, Japan
[2]National Institute of Information and Communications Technology, Japan
[3]Information Technology Center, Nagoya University, Japan

## ABSTRACT

This paper presents a high-intelligibility speech synthesis method for persons with dysarthria caused by athetoid cerebral palsy. The muscular control of such speakers is unstable because of their athetoid symptoms, and their pronunciation is unclear, which makes it difficult for them to communicate. In this paper, we present a method for generating highly intelligible speech that preserves the individuality of dysarthric speakers by combining Transformer-TTS, CycleVAE-VC, and a LPCNet vocoder. Rather than repairing prosody from the dysarthric speech, this method transfers the dysarthric speaker's individuality to the speech of a healthy person generated by TTS synthesis. This task is both important and challenging. From the results of our evaluation experiments, we confirmed that the proposed method can partially transfer the individuality of the target dysarthric speaker while maintaining the intelligibility of the source speech.

***Index Terms*—** dysarthria, speech synthesis, text-to-speech, voice conversion, neural vocoder

## 1. INTRODUCTION

This paper focuses on persons with dysarthria caused by athetoid cerebral palsy. These people are prone to frequent involuntary muscle movements and concomitantly unstable speech movements. Their speech tends to be unnatural and unintelligible, which is a great hindrance to their being able to take part in social activities. This is why there is a great need for a speech synthesis method to aid them in their communication. To improve the intelligibility of dysarthric speech, various methods have been proposed and they can be divided into two categories: enhancing the intelligibility of dysarthric speech using voice conversion techniques and building a TTS system that can synthesize high-intelligibility speech. In this paper, we focus on the latter approach. In this approach, it should be noted that preserving individuality is an essential requirement because many people with dysarthria want to communicate using their own voice.

Text-to-speech (TTS) is one of the important technologies for speech communication, and it has long been a subject of research. Recently, instead of approaches based on the Hidden Markov Model [1], various approaches based on deep neural network (DNN) have been proposed. Along with the development of neural vocoders which directly generate audio samples using DNN, these methods can synthesize high-quality speech that is close to natural speech [2].

Voice conversion (VC) is a technique for transforming acoustic domains such as speaker identity, prosody, and emotion while preserving the linguistic information of the speech. Various DNN-based methods have also been proposed to replace conventional Gaussian mixture model (GMM)-based approaches [3] in VC. Particularly, VC methods based on generative models, such as the Variational Autoencoder (VAE) [4] and Generative Adversarial Networks (GAN) [5], have attracted much attention as high-quality VC methods that do not require parallel speech for training [6–8].

Popular approaches that focus on the enhancement of intelligibility include the following: a GMM-based method to convert formant and vowel features [9], building a dictionary based on non-negative matrix factorization (NMF) [10], GAN-based approaches [11, 12], and knowledge distillation from an end-to-end TTS model trained by healthy speech into dysarthric speech recognition model [13]. One concrete application of these approaches is to improve intelligibility in real-time by recognizing the speech of impaired people when they are speaking. However, these approaches are not always appropriate for people with dysarthria due to cerebral palsy, which we focus on in this study, because the speech act itself is a burden to the patient.

In contrast, the TTS-based approach solves the above problems relatively easily by providing input devices that are less burdensome for the patient. As the most popular approach, [14] recorded the speech of patients with amyotrophic lateral sclerosis (ALS) before their speech deteriorates to build personalized TTS systems. However, because the athetoid cerebral palsy is a congenital disease in almost all cases, this approach cannot be applied to this need. What follows are approaches for dealing with the dysarthric speeech caused by athetoid cerebral palsy. In [15], the HMM-based TTS system was proposed that improve intelligibility by applying corrections for duration and fundamental frequency using an unimpaired person's model. [16] proposed the high-intelligibility speech synthesis system that connects a healthy TTS model based on bi-directional long short-term memory and CycleGAN-VC. However, the quality and conversion performance was insufficient due to the deterioration of the quality of the TTS speech and the inadequate ability to ensure the intelligibility of a healthy speaker in CycleGAN-VC.

In this paper, we present a high-intelligibility speech synthesis method for dysarthric speakers using transformer-based TTS [17, 18], CycleVAE-VC [7], and a LPCNet vocoder [19] as state-of-the-art high-fidelity neural TTS, VC, and vocoder. First, a transformer TTS is trained with transcribed unimpaired speech. Second, acoustic features output by TTS are converted by CycleVAE-VC so they have the individuality of the target dysarthric speaker. CycleVAE-VC is one of the non-parallel VC architectures, and has been pro-

[*]Work performed during an internship at NICT.

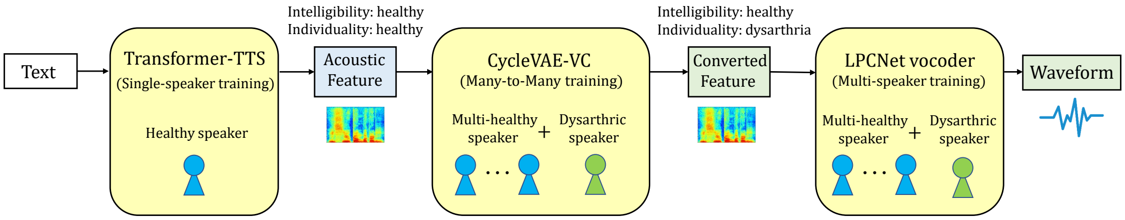

**Fig. 1**. The flow of the proposed method and training dataset on each component.

vided as a baseline model for a cross-lingual voice conversion task in the Voice Conversion Challenge 2020 (VCC2020) [20,21]. We assume that healthy-to-dysarthric VC is a similar task to cross-lingual VC, which transfers speaker identity between domains with different linguistic characteristics, and we expect it can transfer the individuality of dysarthric speech to healthy speech ignoring the collapsed language features of the target dysarthric speaker. Finally, the converted acoustic features are synthesized into speech using a LPCNet vocoder. In [22], we confirmed that LPCNet is capable of sufficient training even in cases where a large amount of speech data cannot be prepared, such as in the case of those with dysarthria, and LPCNet outperforms the Parallel WaveGAN [23] used in a baseline system in [21] that was trained using a small amount of training data.

## 2. PROPOSED METHOD

The architecture of our proposed method is shown in Fig. 1. The proposed method consists of three components: (1) Transformer-TTS is trained using speech data of physically unimpaired person; (2) Non-parallel VC with CycleVAE is performed to give the individuality of the target dysarthric speaker to acoustic features of healthy speech synthesized by TTS; and (3) Converted features are synthesized into speech using a LPCNet vocoder.

### 2.1. Text-to-speech

For TTS, we use the Transformer-based acoustic model which archived state-of-the-art speech quality [18]. This TTS model is trained on the speech data of a Japanese female healthy speaker and produces acoustic features with high intelligibility. The input to this model is not only the phoneme sequence but also the context features for accent estimation. It has been mentioned in [24] that it is necessary to add an external input in addition to the phoneme sequence for speech synthesis in a pitch-accent language such as Japanese. The output of the model is 32-dimensional acoustic features consisting of a 30-dimensional Bark-Frequency Cepstrum Coefficients (BFCCs), pitch period, and pitch correlation. In the following, we will refer to these features as LPCN features.

### 2.2. Non-parallel VC based on Cyclic VAE

For healthy-to-dysarthria VC, this paper use CycleVAE, which is based on VAE-VC with an additional cyclic-consistent approach. In VAE-VC, the encoder compresses the acoustic features of the input into latent space and then the decoder takes the one-hot speaker code and reconstructs the features. By conditioning the speaker code, the latent space is assumed to contain only speaker-independent linguistic features. However, it has been found in [7] that optimization using only the reconstruction of acoustic features does not provide sufficient conversion performance. CycleVAE re-inputs the converted acoustic features into the encoder and adds a constraint that allows the previous acoustic features to be reconstructed, and this process provides higher conversion performance [7].

In this paper, the input acoustic features are the LPCN features described above. The LPCN feature include the pitch period and pitch correlation corresponding to the excitation signal such as the fundamental frequency and the voiced/unvoiced vectors. In a general VC task, excitation features are converted using a linear transformation because they are non-continuous values. In contrast, pitch period and pitch correlation are continuous values calculated by an open-loop cross-correlation search algorithm. Thus, all LPCN features can be subjected to non-linear conversion using CycleVAE.

As mentioned above, since CycleVAE is conditioned on one-hot speaker codes, it is assumed that only linguistic features, such as phonetics, are represented in the latent space. When training with both healthy and dysarthric speech, healthy and collapsed linguistic features will be mapped to the identical linear space. If the latent space properly represents the relationship between healthy and collapsed linguistic features, it should be possible to perform high-quality VC from healthy to dysarthric speakers.

### 2.3. LPCNet

The acoustic features estimated by CycleVAE are synthesized into speech by LPCNet. LPCNet is a WaveRNN-based neural vocoder model with recurrent neural network architecture. LPCNet predicts residual signals between natural speech and predicted speech computed using linear prediction coding (LPC). The original LPCNet was designed to perform speech synthesis with a sampling frequency of 16 kHz, but it has been modified in this paper to perform 24 kHz synthesis. Specifically, as mentioned above, we expand the number of BFCCs from 18 to 30 as [25].

For synthesizing the speech of a dysarthric speaker with improved intelligibility, there are two types of training datasets: only using the speech data of the target dysarthric speaker, and also using the speech of healthy speakers. The former case guarantees the individuality of synthetic speech, but noise is mixed when trying to synthesize speech with improved intelligibility. The latter case suppresses the noise in the synthetic speech, but the individuality is lost to some extent. Additionally, we perform the data augmentation proposed in [21] to obtain a more robust vocoder against mismatches between naturally extracted and converted acoustic features. Specifically, we train LPCNet with a natural feature, a reconstructed feature and a cyclically reconstructed feature as input features. In this paper, the case of using only the target speech of dysarthric speaker refers to single-speaker training (SS), and the case of using speech data of dysarthric and healthy speakers and reconstructed features of CycleVAE refers to multi-speaker training (MS).

## 3. EXPERIMENT

### 3.1. Experimental conditions

**Dataset:** As the speech data of the healthy subjects, we used the single-Japanese female speech database in the JSUT corpus [26] and the multi-Japanese speech database in the JVS corpus [26]. As the speech data of the dysarthric subject, we recorded speech uttered by one subject having dysarthria caused by athetoid cerebral palsy. The dysarthric subject read 430 sentences included in the ATR Japanese speech database [27]. Also, for all the utterances, the sampling frequency is adjusted to 24 kHz.

**Evaluation conditions:** As the baseline, we evaluated the CycleVAE using the WORLD [28] features with Parallel WaveGAN [23] provided as a baseline system in VCC2020 (CycVAE_PWG), and non-cyclic VAE using LPCN features (VAE_LPCN). As the proposed method, we evaluated two conditions of SS and MS training on LPCNet (CycVAE_LPCN). Additionally, as the reference speech for individuality and intelligibility evaluations, we used WSOLA [29] to convert the speaking rate of the dysarthric speech to healthy speakers. This is to reduce the difficulty of evaluation due to different speaking rates. Also, only in the phoneme error rate (PER) experiment described later, the analysis synthesis (AS) of the dysarthric subject was included in the condition.

**Acoustic model:** For the TTS conditions, the Transformer-based acoustic model was trained using 4,800 sentences in the JSUT corpus (Basic5000-0201 to Basic5000-5000) because HTS-style context labels based on manual annotation are available.[1] 100 utterances (Basic5000-0101 to Basic5000-0200) were used for validation and the remaining 100 utterances were used for evaluation. Simple 47-dimensional vectors constructed from 38-dimensional phoneme one-hot vectors and 9-dimensional accentual label vectors were used for the acoustic models. The network architecture was based on Transformer-TTS in ESPNet-TTS [30] implementation and we conducted some modifications to input the accent label vectors. The output features of the acoustic model for (Cyc)VAE_LPCN and CycVAE_PWG were 32-dimensional LPCN features and 55-dimensional WORLD features (voice/unvoiced vector, continuous logF0, 3-dimensional aperiodicity components and 50-dimensional mel-cepstra) with a frame shift of 10 ms, respectively.

**Voice conversion:** For VC conditions, CycleVAE using WORLD features (CycVAE_PWG), VAE using LPCN features (VAE_LPCN) and CycleVAE using LPCN features (CycVAE_LPCN) were used. All implementations were based on the official implementation of VCC2020 and we conducted some modifications for each condition. For training, we used 100 utterances in the JSUT courpus (Basic5000-0001 to Basic5000-0100), 100 utterances generated by TTS (Basic5000-0001 to Basic5000-0100), 100 utterances by four male speakers in the JVS courpus (parallel100 of JVS001, JVS003, JVS005, and JVS006) and 100 utterances by the dysarthric speaker. Although the utterances of the JVS speakers were not used for evaluation, they were used in training to assist in proper optimization of VAE. For evaluation, we used 20 sentences synthesized by TTS included in the ATR Japanese speech database.

**Neural vocoder:** We used LPCNet or Parallel WaveGAN according to each experimental condition. The network architecture of LPCNet was based on the official implementation [19] and we conducted modifications for expansion of the number of BFCCs dimensions. The network architecture of Parallel WaveGAN was the same as an implementation in [21]. For SS training, we used 370 utterances by the dysarthric subject. For MS training, in addition to the datasets in SS training, we use a total of 3000 non-parallel speech samples uttered by 100 Japanese subjects in the JVS corpus and pairs of reconstructed features estimated by CycleVAE and their correct speech (training dataset of JSUT, JVS001, JVS003, JVS005 and JVS006 under VC conditions). As a result of preliminary experiments, the quality deteriorated when the reconstructed features of the dysarthric subject were used for training, so these were not used.

**Table 1**. Results of Naturalness MOS test with 10 listening subjects.

| Method | Score | Method | Score |
|---|---|---|---|
| ORIGINAL | $4.15 \pm 0.16$ | CycVAE_PWG_MS | $1.79 \pm 0.12$ |
| WSOLA | $3.72 \pm 0.15$ | VAE_LPCN_SS | $1.98 \pm 0.17$ |
| JSUT-TTS | $4.69 \pm 0.07$ | CycVAE_LPCN_SS | $2.39 \pm 0.11$ |
| CycVAE_PWG_SS | $1.68 \pm 0.11$ | CycVAE_LPCN_MS | $\mathbf{2.71 \pm 0.11}$ |

### 3.2. Naturalness evaluation

We conducted mean opinion score (MOS) tests to evaluate the subjective perceptual quality of the synthesized speech waveforms (it was not included intelligibility in the evaluation criteria). Ten native Japanese speakers without hearing loss listened to the synthesized speech samples using headphones (20 utterances × 8 conditions = 160 utterances). AS conditions of the dysarthric subject by LPCN and PWG were not included in the experimental conditions of this study, as the quality of the sound was sufficiently close to the original sound through preliminary experiments.

Table 1 shows the result of the naturalness MOS test. Compared to the original sound and WSOLA, none of the VC conditions could achieve sufficient quality. When comparing the quality of the various VC conditions, all LPCN conditions scored better than the PWG conditions. Since LPCNet is an autoregressive model and can refer to past samples to estimate the current sample, it is more robust to the degradation caused by VC than PWG, which can only refer to the current acoustic features to estimate the waveform. Regarding the LPCN condition, the VAE condition did not achieve a sufficient score because it did not use cyclic optimization. In comparison to MS and SS, MS achieved a higher score. This confirms the superiority of using healthy speech and the reconstruction features of CycleVAE for the training of the neural vocoder.

### 3.3. Individuality evaluation

We conducted an individuality ABX test to make sure that the synthesized speech of the proposed method contains the individuality of the target dysarthric speaker. Subjects were the same as those in the Naturalness experiment. Subjects listened to two randomly arranged reference samples (WSOLA and JSUT-TTS) and a synthetic speech, and rated the speaker's proximity to one of the reference speech on a two-point scale (20 utterances × 3 conditions = 60 utterances).

Table 2 shows the result of the individuality ABX test. CycVAE_PWG_MS showed a significant difference on the WSOLA side. This result is good in itself, but there are still problems in terms of naturalness. VAE_LPCN_MS showed a significant difference on the JSUT-TTS side. This result shows that the lack of cyclic optimization resulted in insufficient normalization of speaker individuality over the latent space. CycleVAE_LPCN_MS was not significantly

[1] https://github.com/sarulab-speech/jsut-label

**Table 2**. Results of individuality ABX test with 10 listening subjects.

| Method | WSOLA | JSUT-TTS | $p$-value |
|---|---|---|---|
| CycVAE_PWG_MS | 0.67 | 0.33 | $2.1 \times 10^{-6}$ |
| VAE_LPCN_SS | 0.21 | 0.79 | $3.8 \times 10^{-16}$ |
| CycVAE_LPCN_MS | 0.49 | 0.51 | $7.0 \times 10^{-2}$ |

**Table 3**. Results of intelligibility AB test with 10 listening subjects.

| Method | Score | $p$-value | Conventional |
|---|---|---|---|
| CycVAE_<br>PWG_MS | 0.50 vs. 0.33 | $7.9 \times 10^{-1}$ | WSOLA |
| CycVAE_<br>LPCN_MS | **0.63** vs. 0.22 | $4.6 \times 10^{-3}$ | WSOLA |
| CycVAE_<br>LPCN_MS | **0.72** vs. 0.13 | $1.7 \times 10^{-5}$ | CycVAE_<br>PWG_MS |

**Table 4**. PERs [%] for a variety of methods.

| Method | PER | Method | PER |
|---|---|---|---|
| ORIGINAL | 72.6 | CycVAE_PWG_SS | 60.4 |
| WSOLA | 68.5 | CycVAE_PWG_MS | 51.4 |
| PWG_AS_MS | 74.4 | VAE_LPCN_SS | 28.6 |
| LPCN_AS_MS | 75.9 | CycVAE_LPCN_SS | 31.5 |
| JSUT-TTS | 11.8 | CycVAE_LPCN_MS | 36.6 |

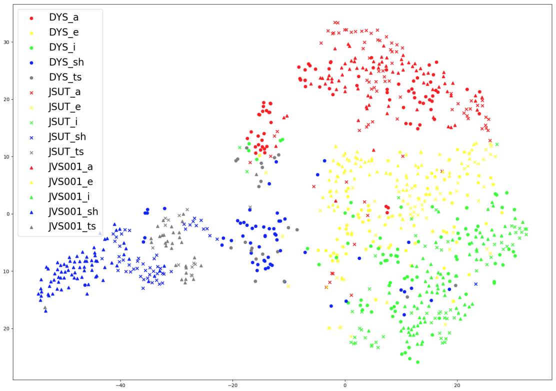

**Fig. 2**. Latent features in three speakers and five different phonemes. The 32-dimensional latent features are compressed and plotted in 2-dimensions using the $t$-SNE algorithm.

different between the two conditions. Subjects commented that they had the individuality of a dysarthric speaker in terms of pitch, but lost some of the individuality due to improvements in prosody and intelligibility. Therefore, although individuality transfer is good, its accurate evaluation of individuality is a challenging task.

### 3.4. Intelligibility evaluation

We conducted a subjective evaluation by AB test and an objective evaluation by phoneme error rate (PER) to evaluate the intelligibility of the proposed method. In the subjective evaluation, subjects were the same as those in the naturalness and individuality experiments. Subjects listened to two randomly arranged samples and rated which one had better intelligibility using the following answers: 1) the former, 2) the latter, and 3) neutral (20 utterances × 3 conditions = 60 utterances). For the objective evaluation, the automatic speech recognition (ASR) system was trained using speech data of multiple unimpaired speakers in the ATR Japanese speech database [27].

Table 3 shows the result of the intelligibility AB test. There was no significant difference between CycVAE_PWG_MS and WSOLA in intelligibility. This is because the quality of the synthesized speech degraded significantly with CycVAE_PWG_MS. This result indicates that the individuality result in section 3.3 was due to a smaller loss of individuality resulting from the improvement in intelligibility. CycVAE_LPCN_MS achieved significant scores against both WSOLA and CycVAE_PWG_MS.

Table 4 shows the PERs of all conditions. The AS conditions have achieved a naturalness close to the original in preliminary experiments, but the score of PER is similar to the original as well. In the VC condition, the VAE_LPCN_SS was closest to the JSUT-TTS score. According to the results of section 3.3, since the VAE_LPCN_SS did not transfer individuality sufficiently, the intelligibility of the VAE_LPCN_SS did not deteriorate much. The proposed method (CycVAE_LPCN) achieved the next highest score, but when comparing SS and MS, the former gave better results. The CycVAE_LPCN_MS sample sometimes contains impulse-like noise that SS does not have, and we assume that it caused the decrease in PER. A detailed investigation is a topic for future work.

### 3.5. Analysis on latent features

We analyzed how the linguistic features of a dysarthric speaker and healthy speaker are distributed in latent space in the CycleVAE. Fig. 2 plots the latent features of five types of Japanese phonemes (a, e, i, sh, and ts) of the two healthy subjects (JSUT and JVS001) and dysarthric speaker (DYS), compressed in two dimensions. The five phonemes are selected as vowels (a, e, and i), which are relatively easy for dysarthric speakers to pronounce, and unvoiced consonants (sh and ts), which are more difficult to pronounce.

As for the distribution of vowels, although the dispersion was larger in the dysarthric subject than in the healthy subjects, the latent features were distributed in almost the same locations. Therefore, these phonemes can be converted in the same way as the VC between healthy subjects. Regarding unvoiced consonants, the latent features of the two healthy subjects were distributed at almost the same location, while those of the dysarthric subject were distributed at a different location. It means that the CycleVAE was trained in these phonemes as “Healthy” and “Dysarthric” separately. When VC is conducted on these phonemes, the acoustic features when the dysarthric speaker utters “Healthy” phonemes are estimated to maintain high-intelligibility.

## 4. CONCLUSION

In this paper, we presented a high-intelligibility speech synthesis method for dysarthric speakers that is achieved by connecting Transformer-TTS, CycleVAE-VC, and LPCNet. Through experiments, we found that CycleVAE can properly represent the relationship of lingual characteristics between healthy and dysarthric speakers and can perform the advanced VC task. Additionally, we also found that LPCNet worked robustly against features that were degraded by VC. In future work, we will consider a method to improve the naturalness of the converted speech and compare it with other non-parallel VC methods.

## 5. REFERENCES

[1] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proc. IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[2] J. Shen, R. Pang, Ron J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z Chen, Y. Zhang, Y. Wang, RJ Skerry-Ryan, Rif A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavanet on mel spectrogram predictions," in *Proc. ICASSP*, Apr. 2018, pp. 4749–4783.

[3] T. Toda, A. 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, Nov. 2007.

[4] D. P. Kingma and J. Ba, "Auto-encoding variational bayes," in *Proc. ICLR*, Apr. 2014.

[5] I. Goodfellow, J. P-Abadie, M. Mirza, D. W-Farley B. Xu, S. Ozair, A. Courville, and Y. Bengio, "Generative adversaria nets," in *Proc. NIPS*, Dec. 2014, pp. 2672–2680.

[6] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: Nonparallel voice conversion with auxiliary classifier variational autoencoder," *IEEE/ACM Trans. Audio Speech Lang. Process*, vol. 27, no. 9, pp. 1432–1443, Sept. 2019.

[7] P. L. Tobing, Y-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Non-parallel voice conversion with cyclic variational atoencoder," in *Proc. Interspeech*, Sept. 2019, pp. 674–678.

[8] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2:improved CycleGAN-based non-parallel voice conversion," in *Proc. ICASSP*, May 2019, pp. 6820–6824.

[9] 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, Sept. 2007.

[10] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Individuality-preserving voice conversion for articulation disorders based on non-negative matrix factorization," in *Proc. ICASSP*, May 2013, pp. 8037–8040.

[11] Chen L.-W, H.-Y Lee, and Y. Tsao, "Generative adversarial networks for unpaired voice transformation on impaired speech," in *Proc Interspeech*, Sept. 2019, pp. 719–723.

[12] S. H. Yang and M. Chung, "Improving dysarthric speech intelligibility using cycle-consistent adversarial training," in *Proc BIOSIGNALS*, Feb. 2020.

[13] D. Wang, J Yu, X. Wu, S. Liu, L. Sun, X. Liu, and H. Meng, "End-to-end voice conversion via closs-modal knowledge distillation for dysarthric speech reconstruction," in *Proc. ICASSP*, May 2020, pp. 7744–7748.

[14] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities:voice banking and reconstruction," *Acoust. Sci. Tech.*, vol. 33, no. 1, pp. 1–5, Jan. 2012.

[15] R. Ueda, T. Takiguchi, and Y. Ariki, "Individuality-preserving voice reconstruction for articulation disorders using text-to-speech synthesis," in *Proc. ACM ICML*, Nov. 2015, pp. 343–346.

[16] R. Nanzaka and T. Takiguchi, "Hybrid text-to-speech for articulation disorders with a small amount of non-parallel data," in *Proc. APSIPA*, Nov. 2018, pp. 1761–1765.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, . Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, Dec. 2017, pp. 5998–6008.

[18] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. T. Zhou, "Neural speech synthesis with transformer network," in *Proc. AAAI*, Jan. 2019, pp. 6706–6713.

[19] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. ICASSP*, May 2019, pp. 5826–7830.

[20] Y. Zhao, W.-C Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice Conversion Challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, Oct. 2020, pp. 80–98.

[21] P. L. Tobing, Y.-C Wu, and T. Toda, "Baseline system of Voice Conversion Challenge 2020 with cyclic variational autoencoder and Parallel WaveGAN," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, Oct. 2020, pp. 155–159.

[22] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, Y. Shiga, and H. Kawai, "Investigation of training data size for real-time neural vocoders on CPUs," *Acoust. Sci. Tech.*, vol. 42, pp. 65–68, Jan. 2021.

[23] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, May 2020, pp. 6199–6203.

[24] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, "Investigation of enhanced tacotron text-to-speech synthesis systems with self-attention for pitch accent language," in *Proc. ICASSP*, May 2019, pp. 6905–6909.

[25] Y. Zheng, X. Li, F. Xie, and L. Lu, "Improving end-to-end speech synthesis with local recurrent neural network enhanced Transformer," in *Proc. ICASSP*, May 2020, pp. 6734–6738.

[26] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JSUT and JVS: free Japanese voice corpora for accelerating speech synthesis research," *Acoust. Sci. Tech.*, vol. 41, no. 5, pp. 761–768, Sept. 2020.

[27] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Commun.*, vol. 9, no. 4, pp. 357–363, Aug. 1990.

[28] 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, July 2016.

[29] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. ICASSP*, Apr. 1993, pp. 554–557.

[30] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *Proc. ICASSP*, May 2020, pp. 7654–7658.