A Preliminary Study of a Two-Stage Paradigm for Preserving Speaker Identity in Dysarthric Voice Conversion

Wen-Chin Huang[1]*, Kazuhiro Kobayashi*[1]*, Yu-Huai Peng*[2]*, Ching-Feng Liu*[3]*,*
Yu Tsao[2]*, Hsin-Min Wang*[2]*, Tomoki Toda*[1]

[1]Nagoya University, Japan [2]Academia Sinica, Taiwan [3]Chi Mei Hospital

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

Abstract

We propose a new paradigm for maintaining speaker identity in dysarthric voice conversion (DVC). The poor quality of dysarthric speech can be greatly improved by statistical VC, but as the normal speech utterances of a dysarthria patient are nearly impossible to collect, previous work failed to recover the individuality of the patient. In light of this, we suggest a novel, two-stage approach for DVC, which is highly flexible in that no normal speech of the patient is required. First, a powerful parallel sequence-to-sequence model converts the input dysarthric speech into a normal speech of a reference speaker as an intermediate product, and a nonparallel, frame-wise VC model realized with a variational autoencoder then converts the speaker identity of the reference speech back to that of the patient while assumed to be capable of preserving the enhanced quality. We investigate several design options. Experimental evaluation results demonstrate the potential of our approach to improving the quality of the dysarthric speech while maintaining the speaker identity.

Index Terms: dysarthric voice conversion, sequence-to-sequence modeling, nonparallel voice conversion, variational autoencoder

1. Introduction

Dysarthria refers to a type of speech disorder caused by disruptions in the neuromotor interface such as cerebral palsy or amyotrophic lateral sclerosis [1]. Dysarthria patients lack normal control of the primary vocal articulators, resulting in abnormal and unintelligible speech with phoneme loss, unstable prosody, and imprecise articulation. The ability to communicate with speech in everyday life is therefore degraded, and it is of urgent need to improve the intelligibility of the distorted dysarthric speech.[1]

Voice conversion (VC), a technique that aims to convert the speech from a source to that of a target without changing the linguistic content [2], has been a dominant approach for dysarthric speech enhancement. We hereafter refer to this task as DVC. Rule-based transformation based on signal processing [3] is limited in that each patient needs to be individually considered. Statistical approaches adopt models ranging from Gaussian mixture models [4], exemplar-based methods [5, 6] and deep neural networks [7, 8, 9].

One of the most difficult problems in not only DVC but VC for other disordered speech such as alaryngeal speech [10] is how to maintain the patient identity after conversion. This is mainly because collecting normal speech of the patient is almost impossible. There have been attempts to tackle this problem. A one-to-many VC system based on eigenvoice conversion was proposed for alaryngeal speech enhancement, whose setting was still considered too idealized since they assumed that a few normal samples of the patient can still be accessed [11].

Our goal in this work is to utilize VC techniques to convert the patient's dysarthric speech into a more intelligible, more natural speech while maintaining the speaker identity of the patient. In light of this, we propose a novel, two-stage approach that combines recent advances in the field of VC. Figure 1 depicts the general idea of the proposed method. In the first stage, a sequence-to-sequence (seq2seq) model converts the input dysarthric speech into that of a reference normal speaker, where we adopted a Transformer-based model named Voice Transformer Network (VTN) [12]. The ability of seq2seq VC models to convert suprasegmental information and the parallel training strategy can greatly improve the naturalness and intelligibility, though the speaker identity is changed into that of the reference speaker. Next, a frame-wise, nonparallel VC model realized by a variational autoencoder (VAE) [13, 14, 15] takes the converted speech with the identity of the reference speaker as input and restores the identity of the patient. An important assumption we make here is that due to the frame-wise constraint, the VAE model changes only time-invariant characteristics such as the speaker identity, while preserving time-variant characteristics, such as pronunciation. As a result, the converted speech has the speaker identity of the patient while maintaining high intelligibility and naturalness. We acknowledge that recently a very similar idea was proposed for preserving speaker identity in not DVC but dysarthric TTS [16].

We evaluate our proposed method on a Mandarin corpus collected from a female cerebral palsy patient. We investigate the importance of the reference speaker choice, and examine how much the aforementioned assumption holds with the current VAE model we adopt. Finally, objective and subjective evaluations show that our approach can improve the naturalness and intelligibility of the dysarthric speech.

Our main contributions in this work are as follows:

- We show that the proposed two-stage method for DVC can restore the patient identity without any normal speech of the patient while improving naturalness and intelligibility.
- To our knowledge, this is the first work to evaluate seq2seq modeling for DVC with a more complex dataset rather than single-worded datasets [8].

2. Related works

2.1. Sequence-to-sequence voice conversion

Compared with conventional frame-wise VC methods [2, 17], seq2seq VC has shown its extraordinary ability in converting

[1]In the field of VC, orthogonal descriptions such as "naturalness" and "intelligibility" are often used, but we use the term "quality" in this paper interchangeably.

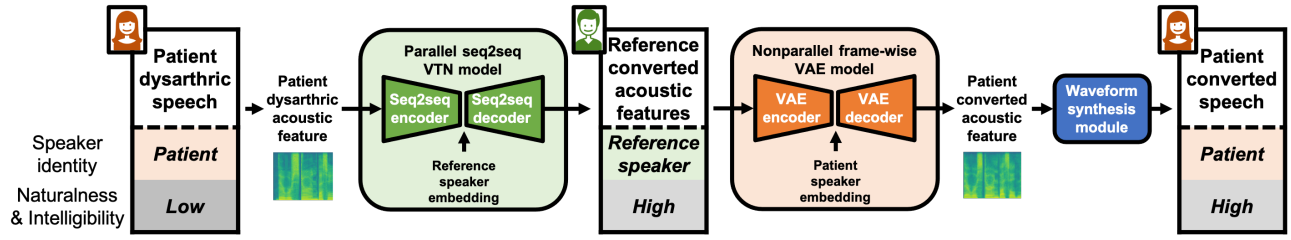


Figure 1: *Illustration of the conversion process in the proposed two-stage method for preserving speaker identity in dysarthric voice conversion.*

speaker identity [12, 18, 19, 20, 21]. Seq2seq modeling is capable of generating outputs of various lengths and capturing long-term dependencies [22], making it a suitable choice for handling suprasegmental characteristics in speech including F0 and duration patterns. However, due to parallel training, applying seq2seq modeling to DVC results in unwanted change of speaker identity.

2.2. Nonparallel frame-wise voice conversion based on variational autoencoder

Nonparallel VC is attractive in that no parallel utterances between the source and target are required, and one of the major trends is autoencoder-based methods [14, 23, 24, 25, 26]. Specifically, the encoder first encodes the input feature into a latent code, and the decoder then mixes the latent code and a specified target speaker embedding to generate the converted feature. Autoencoders are usually trained with a reconstruction loss, but many techniques have been applied to solve various problems in training. The use of a variational autoencoder (VAE) [13] is the most widely adopted method since it greatly stablizes training [14, 23, 26]. Other techniques include using generative adversarial networks (GANs) to alleviate oversmoothing [23, 24, 26], introducing a cyclic objective to improve conversion performance [25], or applying vector-quantization [27] which introduces discreteness into the latent space to capture the categorial property of the linguistic contents in speech.

VAE-based VC is categorized into the frame-based method, which tends not to convert supra-segmental features very well. Although the conversion similarity is therefore inferior to seq2seq-based methods, there are applications where it would be better to keep them unchanged, such as cross-linugal VC and, DVC.

3. Proposed method

An overview of our proposed method is illustrated in Figure 1. Assume that we have a parallel corpus between the dysarthria patient and multiple reference speakers. Our proposed method consists of two models: a seq2seq model that converts the acoustic feature sequence extracted from the input dysarthric speech into that of a reference normal speaker to be more intelligible and natural, and a nonparallel frame-wise model that restores the identity of the patient, which is realized by a VAE. To generate the converted waveform from the acoustic feature, we used the parallel waveGAN (PWG) neural vocoder [28] as the waveform synthesis module, which enables parallel, real-time waveform generation. Note that we do not perform waveform generation between the two models. In the following sections we explain details and design choices of the respective modules.

3.1. Sequence-to-sequence modeling

We adopted the many-to-many Voice Transformer Network (VTN) with text-to-speech (TTS) pretraining. VTN is a seq2seq model for VC based on the Transformer model [29], which relies on multi-head self-attention layers to efficiently capture local and global dependencies. It takes acoustic features (e.g. log mel spectrograms) as input and outputs converted acoustic features. It was extended to a many-to-many version in [30], which was shown to be more effective when a parallel corpus between multiple speakers is available.

The TTS pretraining technique is a two-stage process that transfers the core ability of a seq2seq VC model, which is to encode linguistic-rich hidden representations, from large-scale TTS datasets [12, 31]. First, the decoder pretraining essentially involves training a TTS model on a large-scale TTS dataset. Using the same TTS corpus as input and target, the encoder is then pretrained with a reconstruction loss by fixing the learned decoder from the first stage. Since the decoder was trained to recognize the linguistic-rich hidden representations encoded from text, the encoder hence learns to extract representations of similar properties from speech. The VC model training is finally performed with the VC corpus, which can be completely different from the TTS corpus in terms of speaker and content.

It is worth investigating the choice of the reference speaker. Although the corpus was designed to be parallel among the patient and all reference speakers, due to the difference in characteristics such as the speaking rate and F0 pattern, some speakers can be easier to convert to, compared to others. We thus hypothesize that by choosing a reference speaker with similar characteristics to the patient, conversion might be made easier. We define the similarity between the patient and the reference speaker to be the best performance the VC model can obtain. In later sections, we present our analysis on how the choice of reference speaker affects the conversion performance in various aspects.

3.2. Nonparallel frame-wise model

For the VAE model, we used *crank* [15], an open-source VC software that combines recent advances in autoencoder-based VC methods, including the use of hierarchical architectures, cyclic loss and adversarial training. To take full advantage of unsupervised learning, we trained the network using not only the data of the patient and the reference speakers but also a multi-speaker TTS dataset.

4. Experimental Evaluations

4.1. Experimental settings

To collect the dysarthric speech dataset, a female patient was asked to read the prompts in the phonetically-balanced

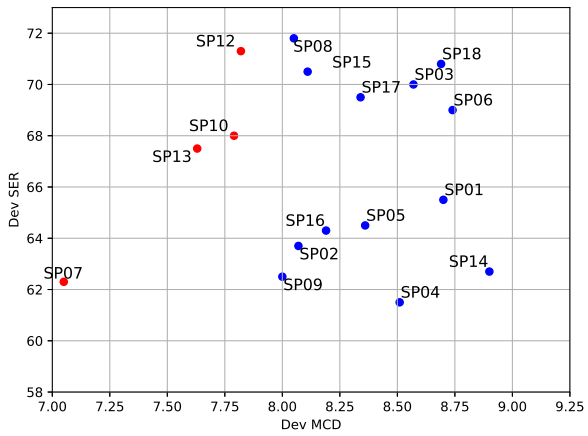


Figure 2: *Scatter plots of the MCD and SER scores of each speaker. Both MCD and SER are the lower the better. Red and blue dots denote female and male speakers, respectively.*

TMHINT dataset [32], where each of the 320 sentences contained 10 Mandarin characters. For the reference speakers, we used the audio recordings of 17 speakers (13 male and 4 female speakers)[2] in the TMSV dataset [33], which was also collected with the TMHINT prompts. We used a 240/40/40 train/validation/test split. All speech utterances were downsampled to 16 kHz, and 80-dimensional mel spectrograms with a 16 ms frame shift were extracted as the acoustic feature.

The implementation of the VTN was based on the open-source toolkit ESPnet [34, 35]. The detailed configuration can be found online[3]. The TTS pretraining was conducted with the Sinica COSPRO multi-speaker Mandarin dataset [36], which is 44 hr long. The implementation of VAE was based on *crank*, which can also be accessed freely[4]. Sinica COSPRO was used along with the TMSV and the patient's voice as training data for the VAE training. For the PWG, we followed an open-source implementation[5]. The training data of PWG contained the audio recordings of the 18 TMSV speakers.

4.2. Objective evaluation

We carried out two types of objective evaluation. First, the mel cepstrum distortion (MCD) is a commonly used measure of spectral distortion in VC, which can only be calculated when the ground truth sample is available. We thus only used this metric in the evaluation of the VTN model. Second, to evaluate the intelligibility of the VC system, we used a Transformer-based automatic speech recognition (ASR) engine pretrained on the AISHELL-1 dataset [37] to transcribe the converted speech, and directly calculated the character error rate (CER) based on the ASR outputs. We then converted the characters into pinyin and discarded the tone to obtain the syllable error rate (SER) of the converted speech.

4.2.1. Investigation of the choice of reference speaker

We first examine our hypothesis on the importance of the choice of reference speaker, as described in Section 3.1. Since we carried out two types of objective metrics, it is worthwhile to examine which is a more proper selection criterion. We trained the VTN model for 2000 epochs, and we selected the best performing models based on MCD. The results are shown in Figure 2.

First, for MCD, female reference speakers tend to yield lower scores, which is reasonable since the patient is also a female. On the other hand, the SER scores did not differ much between genders, and none of the genders gave obviously lower scores. Nonetheless, the speaker with the lowest MCD score (SP07) did not necessarily give the lowest SER value (SP09 gave the lowest SER value), and vice versa. To examine which criterion is better, we conducted a listening test as a more reliable proof, which will be presented in later sections.

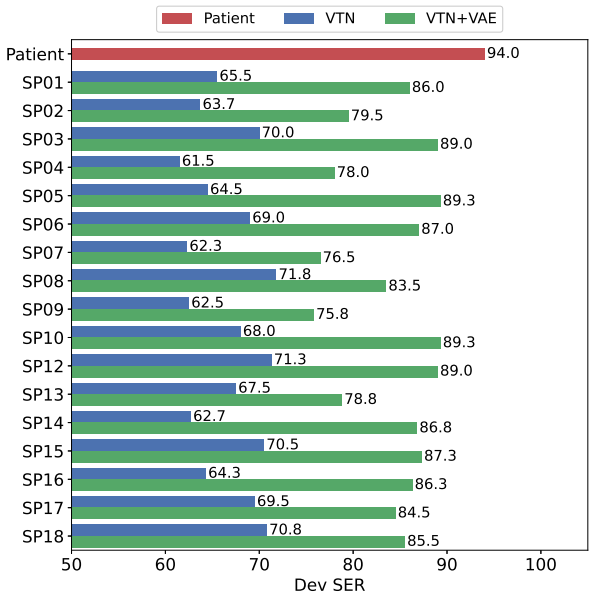


Figure 3: *SER values of the patient's dysarthric voice and the output speech after the VTN model and the VAE model. The values are the smaller the better.*

4.2.2. Intelligibility degradation from VAE

As stated in Section 1, an important assumption in this work is that the intelligibility should be consistent throughout the VAE model. We examine how valid this assumption is by comparing the SERs of the original dysarthric voice as well as the output speech after the VTN model and the VAE model. The results are shown in Figure 3. It can be clearly observed that our assumption was not held, as all SER values after the VAE model are much higher than those of the VTN model output. This is because of insufficient unsupervised factorization in the VAE model we used. As a result, a well-shared linguistic representations space between the normal speech and the dysarthric speech cannot be learned.

Nonetheless, the conversion pairs with most of the reference speakers still yielded lower values compared with the original dysarthric speech. Specifically, speaker SP09 gave the lowest SER of 75.8 after the complete conversion process, which was 18.2 points lower than the original 94.0. This result demonstrates the effectiveness of the proposed two-stage method. In later sections, we will further examine the degradation of naturalness with the listening test results.

[2] Speaker SP11 was excluded due to labeling error.

[3] https://github.com/espnet/espnet/tree/master/egs/arctic/vc1

[4] https://github.com/k2kobayashi/crank

[5] https://github.com/kan-bayashi/ParallelWaveGAN

Table 1: *Results of subjective evaluation using the test set with 95% confidence intervals. All values are higher the better.*

	Naturalness				Similarity			
Description	SP04	SP09	SP07	SP13	SP04	SP09	SP07	SP13
Dysarthric	$2.37 \pm .19$				—			
TMSV	$4.99 \pm .01$				$9\% \pm 7\%$			
VTN	$3.29 \pm .32$	$3.16 \pm .27$	$3.45 \pm .37$	$3.74 \pm .27$	$8\% \pm 8\%$	$8\% \pm 9\%$	$30\% \pm 11\%$	$25\% \pm 14\%$
VTN+VAE	$2.42 \pm .30$	$2.38 \pm .41$	$2.65 \pm .39$	$2.60 \pm .35$	$45\% \pm 10\%$	$45\% \pm 14\%$	$49\% \pm 11\%$	$42\% \pm 11\%$

4.3. Subjective evaluation

We conducted subjective tests on naturalness and conversion similarity to evaluate the perceptual performance. Since it is impractical to evaluate all converted samples of the 17 reference speakers, for both metrics we chose two speakers with the lowest values (MCD: SP07, SP13; SER: SP04, SP09). For naturalness, participants were asked to evaluate the naturalness of the speech by the mean opinion score (MOS) test on a five-point scale. For conversion similarity, each listener was presented a natural target speech and a converted speech, and asked to judge whether they were produced by the same speaker on a four-point scale (Definitely the same, the same, different, definitely different). We recruited 11 native Mandarin speakers. Table 1 shows the results. Audio samples are available online[6].

4.3.1. Investigation of the choice of reference speaker

We first continued our investigation on the reference speaker. For naturalness, as expected, the reference speakers with lower MCD values (SP07, SP13) outperformed the other two speakers (Sp04, SP09). Surprisingly, even after the VAE conversion, SP07 and SP13 still yielded better performances. This shows that listeners payed less attention to the intelligibility, but valued other factors such as fluency and stability more. This also explains why the dysarthric speech, although with extremely low intelligibility, still yielded a MOS score of 2.37. On the other hand, for similarity, such trend was not so obvious, as only SP07 slightly outperformed the other two speakers, and the difference was not significant. Overall, the best performing reference speaker was SP07, whose naturalness (2.65) and similarity (49%) scores were the best among all other speakers after VTN and VAE.

4.3.2. Naturalness degradation from VAE

Next, we continued to examine the naturalness consistency assumption described in Section 1. it could be clearly observed that, regardless of which reference speaker, the naturalness scores degraded for almost 1 MOS point, showing that the current VAE model could not guarantee such consistency, which is similar to the findings in Section 4.2.2. Nonetheless, the best performing speaker, SP07, yielded a naturalness MOS of 2.65, which was shown to be significantly better than 2.37, the MOS given by the dysarthric speech. This result again demonstrated the effectiveness of the proposed method.

4.3.3. Identity preservation ability

We finally examined the ability of our proposed method in maintaining speaker identity. Although the best similarity score of our method could achieve was only 49%, feedbacks from the listeners suggested that it was easy to find the converted speech different from that of the dysarthric speech due to its special characteristics. Since the normal speech of the patient is impossible to obtain, it is essentially difficult to evaluate conversion similarity. To this end, we concluded that the result was acceptable in this preliminary study, and would like to leave the improvement as future work.

[6] `https://bit.ly/3sHxaGY`

5. Conclusions and Discussions

In this paper, we proposed a novel two-stage paradigm for maintaining speaker identity in DVC, where a parallel seq2seq model first converts the source dysarthric speech into that of a reference speaker with the quality enhanced, and a nonparallel frame-wise model realized by a VAE then converts the speaker identity back to the patient while preserving the quality. The experimental results showed that (1) the reference speaker with lower MCD is considered better, (2) the current VAE model does not guarantee quality consistency, and (3) our method can still improve the quality to a certain extent while preserving speaker identity. Yet, the current performance is still far from satisfactory, and below we discuss several improving directions.

Improve seq2seq modeling. The current intelligibility after the first seq2seq conversion stage was much worse than that of past DVC works on simpler datasets [8]. Although we believe this is due to the more complex, limited dataset we used, it is worthwhile to apply techniques like text supervision [20] or data augmentation [21].

Quality consistency assumption. The current VAE model we employed could not guarantee to preserve the enhanced quality. Possible directions include a hierarchical structure to modify solely time-invariant characteristics, or resorting to other frame-wise models such as PPG-based methods [38].

Automatic reference speaker selection. In this work, we chose the best reference speaker by examining the MCD and SER scores of the converted speech, which was an ad-hoc approach. To quickly decide the suitable reference speaker for an arbitrary patient, we may further use pretrained speaker representations like x-vectors [39] as a measurement.

6. Acknowledgements

We would like to thank Dr. Hirokazu Kameoka from NTT CS Laboratory, Japan for the fruitful discussions. This work was supported by JST CREST Grant Number JPMJCR19A3, Japan.

7. References

[1] R. D. Kent, "Research on speech motor control and its disorders: A review and prospective," *Journal of Communication disorders*, vol. 33, no. 5, pp. 391–428, 2000.

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

[3] F. Rudzicz, "Adjusting dysarthric speech signals to be more intelligible," *Computer Speech and Language*, vol. 27, no. 6, pp. 1163 – 1177, 2013.

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

[5] S. Fu, P. Li, Y. Lai, C. Yang, L. Hsieh, and Y. Tsao, "Joint dictionary learning-based non-negative matrix factorization for voice conversion to improve speech intelligibility after oral surgery," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 11, pp. 2584–2594, 2017.

[6] Y. Zhao, M. Kuruvilla-Dugdale, and M. Song, "Voice conversion for persons with amyotrophic lateral sclerosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2942–2949, 2020.

[7] C.-Y. Chen, W.-Z. Zheng, S.-S. Wang, Y. Tsao, P.-C. Li, and Y.-H. Lai, "Enhancing Intelligibility of Dysarthric Speech Using Gated Convolutional-Based Voice Conversion System," in *Proc. Interspeech*, 2020, pp. 4686–4690.

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

[9] M. Purohit, M. Patel, H. Malaviya, A. Patil, M. Parmar, N. Shah, S. Doshi, and H. A. Patil, "Intelligibility improvement of dysarthric speech using mmse discogan," in *Proc. SPCOM*, 2020, pp. 1–5.

[10] J. M. Christensen and P. E. Dwyer, "Improving alaryngeal speech intelligibility," *Journal of communication disorders*, vol. 23, no. 6, pp. 445–451, 1990.

[11] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "Alaryngeal speech enhancement based on one-to-many eigenvoice conversion," *IEEE/ACM TASLP*, vol. 22, no. 1, pp. 172–183, 2014.

[12] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," in *Proc. Interspeech*, 2020, pp. 4676–4680.

[13] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[14] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. APISPA ASC*, 2016, pp. 1–6.

[15] K. Kobayashi, W.-C. Huang, Y.-C. Wu, P. L. Tobing, T. Hayashi, and T. Toda, "crank: An open-source software for nonparallel voice conversion based on vector-quantized variational autoencoder," in *Proc. ICASSP*, 2021.

[16] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, Y. Shiga, and H. Kawai, "High-intelligibility Speech Synthesis for Dysarthric Speakers with LPCnet-based TTS And CycleVAE-based VC," in *Proc. ICASSP*, 2021.

[17] T. Toda, A. W. Black, and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," *IEEE/ACM TASLP*, vol. 15, no. 8, pp. 2222–2235, 2007.

[18] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "ATTS2S-VC: Sequence-to-sequence Voice Conversion with Attention and Context Preservation Mechanisms," in *Proc. ICASSP*, 2019, pp. 6805–6809.

[19] J. Zhang, Z. Ling, L. Liu, Y. Jiang, and L. Dai, "Sequence-to-Sequence Acoustic Modeling for Voice Conversion," *IEEE/ACM TASLP*, vol. 27, no. 3, pp. 631–644, 2019.

[20] J.-X. Zhang, Z.-H. Ling, Y. Jiang, L.-J. Liu, C. Liang, and L.-R. Dai, "Improving Sequence-to-sequence Voice Conversion by Adding Text-supervision," in *Proc. ICASSP*, 2019, pp. 6785–6789.

[21] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to Hearing-Impaired Speech and Speech Separation," in *Proc. Interspeech*, 2019, pp. 4115–4119.

[22] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in *Proc. NIPS*, 2014, pp. 3104–3112.

[23] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *Proc. Interspeech*, 2017, pp. 3364–3368.

[24] J. chieh Chou, C. chieh Yeh, H. yi Lee, and L. shan Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," in *Proc. Interspeech*, 2018, pp. 501–505.

[25] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Non-Parallel Voice Conversion with Cyclic Variational Autoencoder," in *Proc. Interspeech*, 2019, pp. 674–678.

[26] W.-C. Huang, H. Luo, H.-T. Hwang, C.-C. Lo, Y.-H. Peng, Y. Tsao, and H.-M. Wang, "Unsupervised representation disentanglement using cross domain features and adversarial learning in variational autoencoder based voice conversion," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 4, pp. 468–479, 2020.

[27] A. v. d. Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proc. NIPS*, 2017, pp. 6309–6318.

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

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[30] H. Kameoka, W. C. Huang, K. Tanaka, T. Kaneko, N. Hojo, and T. Toda, "Many-to-many voice transformer network," *IEEE/ACM TASLP*, vol. 29, pp. 656–670, 2021.

[31] W. C. Huang, T. Hayashi, Y. C. Wu, H. Kameoka, and T. Toda, "Pretraining techniques for sequence-to-sequence voice conversion," *IEEE/ACM TASLP*, vol. 29, pp. 745–755, 2021.

[32] M.-W. Huang, "Development of taiwan mandarin hearing in noise test," Master's thesis, Department of speech language pathology and audiology, National Taipei University of Nursing and Health Science, 2005.

[33] S.-Y. Chuang, H.-M. Wang, and Y. Tsao, "Improved lite audio-visual speech enhancement," *arXiv preprint arXiv:2008.13222*, 2020.

[34] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

[35] S. Watanabe, F. Boyer, X. Chang, P. Guo, T. Hayashi, Y. Higuchi, T. Hori, W.-C. Huang, H. Inaguma, N. Kamo *et al.*, "The 2020 espnet update: new features, broadened applications, performance improvements, and future plans," *arXiv preprint arXiv:2012.13006*, 2020.

[36] C.-Y. Tseng, Y.-C. Cheng, and C. Chang, "Sinica cospro and toolkit - corpora and platform of mandarin chinese fluent speech," in *Proc. O-COCOSDA*, 2005, pp. 23–28.

[37] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Proc. O-COCOSDA*, 2017, pp. 1–5.

[38] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. ICME*, 2016, pp. 1–6.

[39] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.