A STUDY ON COMBINING NON-PARALLEL AND PARALLEL METHODOLOGIES FOR MANDARIN-ENGLISH CROSS-LINGUAL VOICE CONVERSION

Chang Huai You, Minghui Dong

Institute for Infocomm Research, A*STAR, Singapore

ABSTRACT

In this paper, we propose a cross-lingual voice conversion (VC) scheme leveraging non-parallel and parallel methodologies. The goal of the VC is to transform the voice of one speaker from a language dataset into the voice of another speaker from a different language dataset. First, two non-parallel methods are separately investigated, they are CyclGAN-VC2 and phonetic posterior-grams (PPG) VC. Second, two different parallel VC systems are developed to enhance the quality of the converted speech spectrogram, where the output speech from the non-parallel VC is used to form the parallel pair with the corresponding original speech. Focusing on Mandarin-English bilingual databases, the proposed VC scheme improves speech naturalness and speaker similarity as compared to the baseline non-parallel methods.
Index Terms: non-parallel voice conversion, parallel voice conversion, generative adversarial network, text-to-speech, phonetic posterior-grams

1. INTRODUCTION

Integrating multilingual text-to-speech (TTS) into a single synthesis model has been studied in many literatures [1] [2] [3]. Cross-lingual voice conversion has been used for multilingual TTS to achieve speech generation [4]. In [5], the zero-shot many-to-many speaker voice conversion (VC) is studied. However, many-to-many VC model degrades the quality of converted voices when many speakers are involved for training simultaneously. The single speaker database is of advantage on high quality bilingual TTS because the interference of variation of different speaker voices is avoided. However, two different language corpora spoken by the same professional speaker is very challenging for recording since the professional speaker is usually fluent for only one language. Cross-lingual VC is a way to unify the voices of different language speakers.

Generally, VC can be categorized into parallel method and non-parallel method. Parallel method relies on the availability of parallel utterance pairs of the source and target speakers. There have been many parallel voice conversion techniques like Gaussian mixture model (GMM) [6] [7] [8] and neural networks (NN) approaches [9] [10]. In traditional parallel VC, acoustic features of parallel utterances needs to be aligned frame-by-frame using alignment algorithms, such as dynamic time warping (DTW). The traditional conversion model is trained to learn the mapping function between time-aligned source and target acoustic features. For the cross-lingual VC, however, there are no pair of source and target utterances for the different language. Therefore, directly applying the parallel VC method is not possible for the cross-lingual situation.

On contrast to traditional parallel VC modeling that requires expensive parallel speech data and time sequence alignment, non-parallel VC does not rely on parallel utterances. This characteristics of non-parallel VC makes itself be suitable for cross-lingual VC. CycleGAN-VC is a typical non-parallel voice conversion (VC) method [11], it employs a sequential mapping approach through the utilization of cycle-consistent generative adversarial network (CycleGAN) [12] [13] [14]. StarGAN-VC provides a promising solution by extending CycleGAN-VC to a conditional setting and incorporating domain codes [15]. CycleGAN-VC2 [16] is an improved version of CycleGAN-VC by incorporating improved aspects on adversarial losses, generator and discriminator by replacing FullGAN [17] with PatchGAN [18] [19]. Recently, phonetic posterior-grams (PPG) was used as feature input to a Tacotron based sequence-to-sequence model for cross-lingual VC [20]. However, the damaging trait of PPG limits its VC performance because PPG feature keeps only phonetic components but discards not only the speaker voice component but also the prosody and emotional components of the original speech utterance.

Typical non-parallel CycleGAN based VC faces an adverse training condition while PPG based all-to-one non-parallel VC losses the prosody and emotional components of the source speech during inference process. To overcome the non-parallel VC drawback and complement the loss components from non-parallel VCs, we propose a cross-lingual VC scheme by leveraging the advantages of non-parallel and parallel VC methodologies. Moreover, we develop a parallel Tacotron-VC, in which the frame-level alignment can be ignored since the sequence-to-sequence architecture adopts location attention function.

The remainder of this paper is organized as follows. In session 2, the proposed VC scheme is described in details. The performance evaluation is reported in section 3. We give the conclusion in session 4.

2. PROPOSED VOICE CONVERSION SCHEME

In this paper, we propose a cross-lingual VC scheme leveraging the advantage of both non-parallel and parallel VC methodologies for the voice's substitution between two different language databases. It includes three parts: (1) Non-parallel VC models are developed with two different language databases to generate voice-converted databases; (2) Parallel VC methods are designed by using the speech pairs between the original language data and the corresponding non-parallel

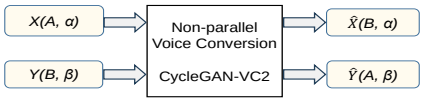

Fig. 1. CycleGAN-VC2 converts the voice of speaker into that of another speaker.

voice-converted data; (3) The parallel VC systems are applied to infer target language databases.

2.1. Condition and Task

Assume that there exists a speech database denoted as $\mathtt{X}(\mathbf{A}, \alpha)$ for language α, which is spoken exclusively by a professional speaker $\mathbf{A}$; alongside another speech database $\mathtt{Y}(\mathbf{B}, \beta)$ for language β, exclusively spoken by another professional speaker $\mathbf{B}$. We want to convert the voice of speech database $\mathtt{X}(\mathbf{A}, \alpha)$ from $\mathbf{A}$ to $\mathbf{B}$, and thus the target database is obtained as $\hat{\mathtt{X}}(\mathbf{B}, \alpha)$.

Vice versa, we can use the similar way to convert the voice of speech database $\mathtt{Y}(\mathbf{B}, \beta)$ from $\mathbf{B}$ to $\mathbf{A}$, and the target database is denoted as $\hat{\mathtt{Y}}(\mathbf{A}, \beta)$.

2.2. Non-parallel VC

Unlike parallel voice conversion, achieving nonparallel voice conversion is typically difficult due to the unfavorable training conditions.

2.2.1. CycleGAN-VC2

Among many nonparallel VC methods [14] [16] [21] [22] [23] [24], CycleGAN-VC2 [16] demonstrates outstanding performance. In this paper, we make use of CycleGAN-VC2 [16] algorithm for the non-parallel conversion. Fig. 1 briefs the non-parallel VC substituting the voice for both databases X and Y. As described in the above session, the data resource for language α is $\mathtt{X}(\mathbf{A}, \alpha)$ and while the resource for language β is $\mathtt{Y}(\mathbf{B}, \beta)$. With two sets of different databases representing language α and language β respectively, we train the CycleGAN-VC2 model to target the converted voice by substituting the voices of speakers A and B of the two databases. As a result, we obtain the target speech databases $\hat{\mathtt{X}}(\mathbf{B}, \alpha)$ and $\hat{\mathtt{Y}}(\mathbf{A}, \beta)$.

2.2.2. PPG-Tacotron VC

We developed another cross-lingual voice conversion utilizing PPG feature. With PPG feature obtained by using automatic speech recognition (ASR) acoustic model, the speaker component can be discarded and only the phonetic posterior probability vectors are kept. When we use the PPG feature to represent the speech data containing only a single voice (we call it as model speaker voice), we can train a non-parallel PPG based VC model. Theoretically, using the PPG VC model, any voice in a speech utterance can be converted into the model speaker voice. The non-parallel PPG VC avoids the parallel voice pair requirements. Different from the parallel VC that is one-to-one conversion, the non-parallel PPG VC is an all-to-one conversion.

To setup a PPG VC system, first, we train ASR acoustic models for PPG feature extraction by using speech recognition database. Second, we select to modify Tacotron architecture [25] for the PPG VC modeling. The model training

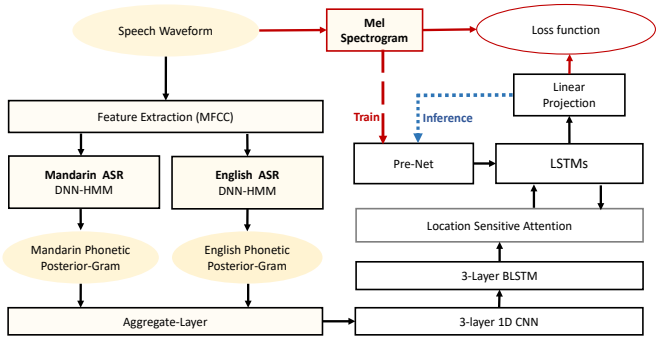

Fig. 2. The PPG-based non-parallel VC by modifying encoder part of Tacotron architecture.

is done by input the PPG features of speech utterances to the modified Tacotron neural network and the spectrograms of the same speech utterances are treated as target of Tacotron output. For instance, we use language database $\mathtt{X}(\mathbf{A}, \alpha)$ to train a PPG VC model, then we convert the voice of another language database $\mathtt{Y}(\mathbf{B}, \beta)$ to the voice of speaker $\mathbf{A}$ to obtain $\hat{\mathtt{Y}}(\mathbf{A}, \beta)$. Fig. 2 illustrates the PPG-Tacotron VC training process in red^{+} as well as inference process in blue^{+}.

We can choose either English ASR model or Mandarin ASR model to extract the PPG feature, however joint PPG feature concatenating the two different ASR acoustic phonetic probabilities gives better performance because the joint PPG feature carries more phonetic details as compared to single ASR PPG.

2.3. Parallel VC Modeling

The CycleGAN method degrades the speech quality due to the disadvantageous training condition. PPG feature ignores the prosody and emotional components of the speech signal while it discards the speaker voice component to preserve only phonetic component. As a result, PPG VC losses source speech prosody and emotional components. To further improve the converted speech quality, we extend the conversion to a parallel VC modeling leveraging the output of the non-parallel VC system (either CycleGAN-VC2 or PPG VC). As a paradigm, the parallel speech is built on the speech pair of the voice-converted speech $\hat{\mathtt{X}}(\mathbf{B}, \alpha)$ and the original speech $\mathtt{X}(\mathbf{A}, \alpha)$. Using voice-converted speech data $\hat{\mathtt{X}}(\mathbf{B}, \alpha)$ as input source of the parallel VC and the original speech data $\mathtt{X}(\mathbf{A}, \alpha)$ as target, the parallel VC neural network model $\mathbf{B2A}$ is trained for converting the voice from $\mathbf{B}$ to $\mathbf{A}$ in spectrogram domain. During this parallel $\mathbf{B2A}$ training, the original $\mathtt{Y}(\mathbf{B}, \beta)$ is also mixed into the model training as input source of the parallel VC and the converted speech data $\hat{\mathtt{Y}}(\mathbf{A}, \beta)$ is used as the corresponding target. This auxiliary data $\mathtt{Y}$ serves solely as a guiding factor for the parallel neural network to see language β. As a result, similar to pre-training, the influence of $\mathtt{Y}$ diminishes gradually as the training iterations progress. Subsequently, the parallel VC model is used to convert the speech data $\mathtt{Y}(\mathbf{B}, \beta)$ to $\hat{\hat{\mathtt{Y}}}(\mathbf{A}, \beta)$. Fig. 3 depicts how to train a parallel VC model for converting the voice of speaker $\mathbf{B}$ into that of speaker $\mathbf{A}$.

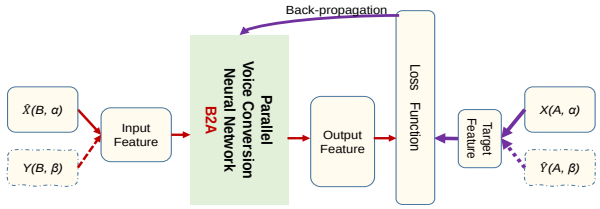

Fig. 3. The **B2A** parallel VC training process converting voice of speaker **B** to voice of speaker **A** by using data `X`, while `Y` is used as auxiliary data.

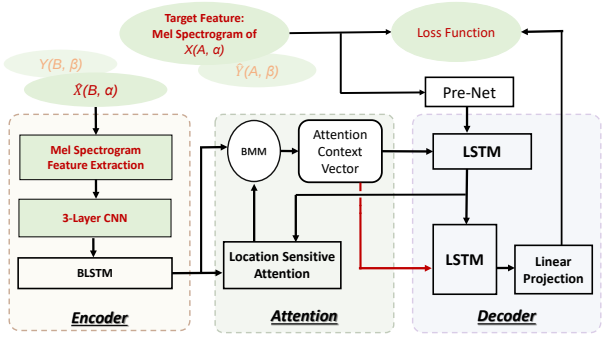

Fig. 4. A paradigm of the parallel VC model training by modifying Tacotron neural network for the conversion of **B2A** using data `X` and auxiliary data `Y`.

2.3.1. *Tacotron Parallel VC:* `Taco-Para`

Tacotron [25] is an open source for speech synthesis combining a sequence-to-sequence recurrent network with location attention to predicts mel spectrograms. We developed a Tacotron parallel VC denoted as `Taco-Para` by modifying Tacotron architecture. Specifically, there are three main parts in Tacotron architecture: encoder, attention, and decoder besides the pre-net and post-net; we remove the text input and vector embedding in encoder part and replace them with the speech signal input and Mel-spectrogram feature extraction. The input to the modified Tacotron parallel VC is the converted speech signal obtained from the non-parallel VC (i.e. CycleGAN-VC2 or PPG-Tacotron VC) system and the target is the 80-dimension spectrogram sequence of the corresponding original speech signal. Fig. 4 depicts the neural network architecture of the modified Tacotron parallel VC, in which the spectrogram is to be converted. As mentioned, during training, the auxiliary data is utilized before fine tune of the parallel model.

2.3.2. *Traditional neural network parallel VC:* `trNN-Para`

Besides Tacotron based parallel VC, we also developed another parallel VC, `trNN-Para`, using traditional neural network architecture without location attention function. To implement `trNN-Para`, we consider the neural network architecture combining convolutional neural network (CNN), bidirectional long short-term memory (BLSTM) network and feed-forward network. In addition, since the `trNN-Para` network lacks of location attention mechanism, we apply the frame level alignment by using DTW for source and target sequences [8].

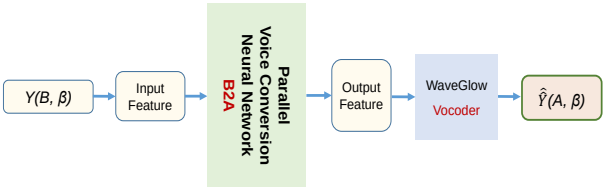

Fig. 5. Inference process: convert voice of English speaker to voice of Mandarin speaker for English database by using the B2A model.

2.4. Target database generation via parallel VC

Following the paradigm in the previous section, the ultimate goal of the parallel VC training is to convert the voice of speaker **B** in database $\mathtt{Y}(\mathbf{B}, \beta)$ into the voice of speaker **A** to obtain the database $\mathtt{Y}(\hat{\mathbf{A}}, \beta)$. Fig. 5 shows the inference flowchat on how to use the above-mentioned parallel VC model (`Taco-Para` VC or `trNN-Para` VC) to convert the voice from **B** into **A** for database `Y`. Finally, output of the inferred spectrogram will go through the Waveglow vocoder [26] to generate the speech signal in time domain. Vice versa, we can use the similar model training and inference method to convert the voice of speaker **A** in database $\mathtt{X}(\mathbf{A}, \alpha)$ into the voice of the speaker **B** to gain the enhanced conversion $\hat{\mathtt{X}}(\mathbf{B}, \alpha)$.

Theoretically, the proposed scheme may improve CycleGAN VC by reusing the defective speech for a parallel VC modeling and the final converted speech is generated from good conditional source speech; it may improve PPG non-parallel VC via the parallel VC inference where the spectrogram feature of source speech carries abundant information including prosody, emotional component, speaker characteristics and phonetic contents as compared to PPG feature that carries only phonetic content.

3. PERFORMANCE EVALUATION

In this section, we conduct the performance evaluation. The databases used for this evaluation are Blizzard10 [27] for Mandarin and LJSpeech [28] for English separately. On the one hand, Blizzard10 database is recorded from an only professional speaker. It contains 5882 utterances, from which 5200 utterances are randomly split for training, 360 utterances are for validation, and the remainder 322 utterances for test. On the other hand, LJSpeech database is recorded from another professional speaker. It contains 13100 utterances, where 12000 utterances are randomly split for training, 600 for validation, and the remainder 500 for test. To explain the experimental conditions, we denote the Blizzard10 database as $\mathtt{X}(\mathbf{A}, \alpha)$ with voice of speaker **A** and language α=Mandarin, and the LJSpeech database as $\mathtt{Y}(\mathbf{B}, \beta)$ with voice of speaker **B** and language β=English. In this evaluation, we use 16 kHz sampling rate.

For non-parallel VC model training, 5000 utterances from Mandarin training database and another 5000 utterances from English training database are selected. For CycleGAN-VC2, the WORLD analyzer [29] is applied to compute 36-dimension Mel-cepstral coefficients (MCEPs), logarithmic fundamental frequency (log F0), and aperiodicities (APs) at 5 ms intervals. During preprocessing, the source and target MCEPs are normalized to zero-mean and unit-variance by using the statistics of the training sets. To train CycleGAN-

Table 1. Mean Opinion Score for cross-lingual voice conversion performance evaluation in terms of naturalness, similarity and intelligibility, where A denotes the voice of Mandarin speaker and B denotes the voice of the English speaker, and `VC-Method1:VC-Method2` denotes the non-parallel VC method-1 followed by parallel VC method-2.

Source	Direction	Voice conversion method	Naturalness	Similarity with target $\mathtt{X}(\mathbf{B},\alpha)$	Intelligibility
Raw Data	A	`Samples from` $\mathtt{X}(\mathbf{A},\alpha)$	4.30±0.12	2.87±0.18	4.29±0.21
	B	`Samples from` $\mathtt{Y}(\mathbf{B},\beta)$	4.32±0.10	4.52±0.10	4.38±0.13
$\mathtt{X}(\mathbf{A},\alpha)$ Mandarin Database	A2B	`CycleGAN-VC2`	3.26±0.22	3.08±0.20	3.28±0.10
		`CycleGAN-VC2:trNN-Para`	3.38±0.16	3.20±0.15	3.40±0.05
		`CycleGAN-VC2:Taco-Para`	3.49±0.26	3.32±0.16	3.47±0.10
		`PPG-Tacotron`	3.74±0.15	3.68±0.20	3.62±0.20
		`PPG-Tacotron:trNN-Para`	3.78±0.10	3.70±0.15	3.72±0.25
		`PPG-Tacotron:Taco-Para`	3.82±0.13	3.79±0.16	3.87±0.21
Language	Direction	Voice conversion method	Naturalness	Similarity with target $\mathtt{Y}(\mathbf{A},\beta)$	Intelligibility
Raw Data	B	`Samples from` $\mathtt{Y}(\mathbf{B},\beta)$	4.32±0.10	2.95±0.15	4.38±0.13
	A	`Samples from` $\mathtt{X}(\mathbf{A},\alpha)$	4.30±0.12	4.50±0.12	4.29±0.21
$\mathtt{Y}(\mathbf{B},\beta)$ English Database	B2A	`CycleGAN-VC2`	3.01±0.16	3.15±0.20	3.20±0.10
		`CycleGAN-VC2:trNN-Para`	3.27±0.12	3.20±0.12	3.36±0.20
		`CycleGAN-VC2:Taco-Para`	3.32±0.22	3.36±0.21	3.48±0.12
		`PPG-Tacotron`	3.82±0.18	3.71±0.16	3.80±0.10
		`PPG-Tacotron:trNN-Para`	3.89±0.10	3.76±0.21	3.85±0.20
		`PPG-Tacotron:Taco-Para`	3.97±0.16	3.80±0.18	3.90±0.12

VC2 model, the Adam optimizer [30] is used and the mini batch size is set to 30 (utterances). The neural network model is trained for 600,000 iterations with learning rates of 0.0002 for the generator and 0.0001 for the discriminator with momentum term of 0.5. During conversion process, the trained CycleGAN-VC2 model is only used to convert MCEPs, while f_0 is converted by using logarithm Gaussian normalized transformation [31] and APs is directly used without modification. Finally, with MCEPs, f_0 and AP as input, the WORLD vocoder [29] synthesizes the speech signal.

In the non-parallel PPG-Tacotron VC, `PPG-Tacotron`, to extract PPG feature, ASR acoustic models for both Mandarin and English are separately trained, where the number of tied-states of triphone indicates the dimension of PPG feature. In our experiments, the dimension of Mandarin PPG feature is 8721 whereas dimension of English PPG is 6344. We choose to report the experimental result with joint PPG feature by concatenating the two ASR phonetic posterior features to form a 15065-dimensional feature. The output layer is to target 80-dimensional spectrograms.

For parallel VC training, we select 2100 pairs from $\hat{\mathtt{Y}}(\mathbf{A},\beta)$ and $\mathtt{Y}(\mathbf{B},\beta)$, and 2100 pairs from both $\hat{\mathtt{X}}(\mathbf{B},\alpha)$ and $\mathtt{X}(\mathbf{A},\alpha)$. Here, $\hat{\mathtt{Y}}(\mathbf{A},\beta)$ and $\hat{\mathtt{X}}(\mathbf{B},\alpha)$ can be from the output of non-parallel VCs such as CycleGAN-VC2 or non-parallel PPG-VC. For both parallel VCs, `Taco-Para` and `trNN-Para`, we use the Adam optimizer with $\beta 1$ = 0.9, $\beta 2$ = 0.999, $\epsilon = 10^{-6}$ and a learning rate of 10^{-3} exponentially decaying to 10^{-5}, the model is trained with 92,000 iterations where the validation loss is converged to stable. In `trNN-Para`, DTW is used to align the feature vector sequence of source and target speech before the parallel VC model training. The specific neural network (NN) architecture is composed of three layers of 1D CNN, two layers of BLSTM with 2048 nodes per layer, and five Layers of feed-forward NN with 1024 nodes each hidden layers and 80 nodes in output layer to target spectrogram. The time-domain speech signal is obtained via WaveGlow vocoder [26] that we have trained for 16 kHz speech generation.

Thirty speech sentences are randomly selected from Mandarin test dataset and English test dataset respectively for subjective evaluation in terms of speech naturalness, speaker similarity and intelligibility[1] based on the rating scale of mean opinion score (MOS), ranging from 1 (poor) to 5 (excellent). [32]. Ten Mandarin native speakers are invited for Mandarin speech naturalness and intelligibility tests and another ten English native speakers for English speech naturalness and intelligibility tests. In addition, ten persons are selected for speaker similarity tests with most of the raters are Mandarin-English bilingual speakers. For speaker similarity, the raters are advised to ignore the speech content but focus only on the speaker identity. Table 1 gives the performance evaluation report by comparing different cross-lingual VC methods. From the experimental result, we can see that our proposed VC scheme improves the performance as compared to the baseline of CycleGAN-VC2 and non-parallel PPG-Tacotron VC methods for both directions from Mandarin speaker to English speaker (A2B) and from English speaker to Mandarin speaker (B2A). Also we can see that the naturalness drops down from the raw speech, which is uttered from the professional speaker, due to the artificial conversion, however, the speaker similarity is relatively improved.

4. CONCLUSION

In this paper, we have proposed a VC scheme leveraging the advantages of non-parallel and parallel VC methodologies for Mandarin-English cross-lingual speech corpora. Through experiments, it is observed that PPG-Tacotron non-parallel VC gives better performance than cycleGAN-VC2 does. It is also noticed that the parallel Tacotron VC (`Taco-Para`) outperforms DTW based traditional neural network approach (`trNN-Para`) for this parallel VC approach. The experimental result shows the effectiveness of the proposed VC scheme in terms of subjective test.

[1]The naturalness pertains to the quality of speech utterance regardless of the content, and the speaker similarity speculates the proximity of the converted voice to the target speaker, whereas the intelligibility assesses the clarity level of the speech content.

5. REFERENCES

[1] C. Traber, K. Huber, K. Nedir, B. Pfister, E. Keller, and B. Zellner, "From multilingual to polyglot speech synthesis," European Conference on Speech Communication and Technology, 1999.

[2] J. Latorre, K. Iwano, and S. Furui, "Polyglot synthesis using a mixture of monolingual corpora," IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., vol. 1. IEEE, 2005, pp. I–1.

[3] Rohan Badlani, Rafael Valle, Kevin J. Shih, Joao Felipe Santos, Siddharth Gururani, Bryan Catanzaro, "Multilingual Multiaccented Multispeaker TTS with RADTTS," *arXiv:2301.10335v1*, NVIDIA, Jan. 2023.

[4] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R.J. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, "Learning to speak fluently in a foreign language:Multilingual speech sy nthesis and cross-language voice cloning," *in Proc. INTERSPEECH*, 2019.

[5] Z. Long, Y. Zheng, M. Yu, and J. Xin, "Enhancing Zero-Shot Many to Many Voice Conversion via Self-Attention VAE with Structurally Regularized Layers," *International Conference on Artificial Intelligence for Industries*, AI4I, 2022.

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

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

[8] K. Kobayashi and T. Toda, "sprocket: Open-Source Voice Conversion Software," *Proc. Odyssey*, pp. 203-210, June 2018.

[9] 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/ACMTrans. Audio Speech Lang. Pr ocess., vol. 22, no. 12, pp. 1859–1872, 2014.

[10] S. Desai, A.W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," IEEE Trans. Audio Speech Lang. Process., vol. 18, no. 5, pp. 954–964, 2010.

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

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," Proc. ICCV, pp. 2223–2232, 2017.

[13] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," Proc. ICCV, pp. 2849–2857, 2017.

[14] T. Kaneko and H. Kameoka "CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks," *EURASIP*, pp. 2114-2118, 2018

[15] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Nonparallel many-to-many voice conversion using star generative adversarial networks," *in Proc. SLT*, 2018, pp. 266–273.

[16] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Im-proved cyclegan-based non-parallel voice conversion," in Proc. ICASSP, 2019, pp. 6820–6824.

[17] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," *in Proc. ICASSP*, pp. 4910–4914, 2017.

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," *in Proc. CVPR*, pp. 5967–5976, 2017.

[19] Chuan Li and Michael Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," *in Proc. ECCV*, pp. 702–716, 2016.

[20] S. Zhao, T. H. Nguyen, H. Wang, and B. Ma, "Towards Natural Bilingual and Code-Switched Speech Synthesis Based on Mix of Monolingual Recordings and Cross-Lingual Voice Conversion," arXiv: 2010.08136v1, Oct. 2020.

[21] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-many voice conversion based on tensor representation of speaker space," Proc. Inter-speech, pp. 653–656, 2011.

[22] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice con-version using variational autoencoders conditioned by phonetic posterior-grams and d-vectors," in Proc. ICASSP, 2018, pp. 5274–5278.

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

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

[25] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in Proc. Interspeech, pp. 4006–4010, Aug. 2017.

[26] Ryan Prenger, Rafael Valle, Bryan Catanzaro, "WAVEGLOW: A FLOW-BASED GENERATIVE NETWORK FOR SPEECH SYNTHESIS," arXiv:1811.00002, https://doi.org/10.48550/arXiv.1811.00002

[27] https://www.synsig.org/index.php, *Blizzard Challenge 2010*

[28] Keith Ito et al., 'The LJ speech dataset," 2017.

[29] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "WORLD: A vocoderbased high-quality speech synthesis system for real-time applications," IEICE Trans. Inf. Syst., vol. 99, no. 7, pp. 1877–1884, 2016.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. ICLR, 2015.

[31] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through phoneme-based linear mapping functions with STRAIGHT for Mandarin," in Proc. FSKD, 2007, pp. 410–414.

[32] "International Telecommunication Union: ITU-T Recommendation P.800.1: Mean Opinion Score (MOS) terminology," Technical report, 2006