IMPROVING RECOGNITION-SYNTHESIS BASED ANY-TO-ONE VOICE CONVERSION WITH CYCLIC TRAINING

Yan-Nian Chen[1]*, Li-Juan Liu*[2]*, Ya-Jun Hu*[2]*, Yuan Jiang*[1,2]*, Zhen-Hua Ling*[1]

[1]National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, P.R.China
[2]iFLYTEK Research, iFLYTEK Co., Ltd., Hefei, P.R.China

yannianchen@mail.ustc.edu.cn, {ljliu,yjhu,yuanjiang}@iflytek.com, zhling@ustc.edu.cn

ABSTRACT

In recognition-synthesis based any-to-one voice conversion (VC), an automatic speech recognition (ASR) model is employed to extract content-related features and a synthesizer is built to predict the acoustic features of the target speaker from the content-related features of any source speakers at the conversion stage. Since source speakers are unknown at the training stage, we have to use the content-related features of the target speaker to estimate the parameters of the synthesizer. This inconsistency between conversion and training stages constrains the speaker similarity of converted speech. To address this issue, a cyclic training method is proposed in this paper. This method designs pseudo-source acoustic features, which are generated by converting the training data of the target speaker towards multiple speakers in a reference corpus. Then, these pseudo-source acoustic features are used as the input of the synthesizer at the training stage to predict the acoustic features of the target speaker and a cyclic reconstruction loss is derived. Experimental results show that our proposed method achieved more consistent accuracy of acoustic feature prediction for various source speakers than the baseline method. It also achieved better similarity of converted speech, especially for the pairs of source and target speakers with distant speaker characteristics.

***Index Terms*—** Voice conversion, any-to-one, recognition-synthesis, cyclic training

1. INTRODUCTION

Voice conversion (VC) aims to convert a speech from a source speaker to make it sound like being spoken by a target speaker, while keeping linguistic contents unchanged [1, 2]. Some VC methods need a parallel training corpus, where the same sentences are spoken by both source and target speakers. The conversion models include Gaussian mixture model (GMM) [3], non-negative matrix factorization (NMF) [4], frequency warping [5], neural networks [6–12] and so on. Considering the difficulty of collecting paired utterances, many non-parallel VC methods without reliance on parallel training data have been explored recently such as recognition-synthesis (Rec-Syn) based ones [13–19], generative adversarial network (GAN) based ones [20, 21], variational auto-encoder (VAE) [18, 22] based ones and disentanglement based ones [23].

Among them, the Rec-Syn based approach has achieved state-of-the-art performance on several VC tasks [15, 19]. In this method, a pre-trained speaker-independent automatic speech recognition (SI-ASR) model is utilized to extract content-related features such as phonetic posteriorgrams (PPGs) or bottleneck features. Then, a synthesizer is trained to predict target acoustic features conditioned on the content-related features. This method assumes that the content-related features contain only speaker-independent linguistic contents. Thus, at the conversion stage, the content-related features extracted from any *source speaker* can be transformed to the acoustic features of the target speaker, i.e., achieving any-to-one VC. However, at the training stage, the content-related features of the *target speaker* are utilized as the input of the synthesizer since source speakers are unknown. Because the content-related features extracted by ASR inevitably contain speaker-dependent information such as timbre and intonation, this inconsistency between training and conversion stages harms the speaker similarity of converted speech, especially for the source speakers whose characteristics are very dissimilar from the target.

To address this issue, some previous studies [24, 25] adopted adversarial learning to reduce the speaker information in content-related features. Zhang *et al.* [25] employed a speaker adversarial loss besides the phoneme recognition loss to train the recognizer. Ding *et al.* [24] applied the speaker adversarial loss to the hidden layer of the synthesizer instead of the recognizer for learning speaker-independent representations.

In this paper, a novel cyclic training method is proposed to deal with this inconsistency issue. In this method, the acoustic features extracted from the training utterances of the target speaker are first fed into an ASR model and an auxiliary synthesizer sequentially to produce pseudo-source acoustic features which are parallel to the training data of the target speaker. Here, the auxiliary synthesizer accepts a speaker code as its input and is trained using a multi-speaker corpus. Then, the pseudo-source acoustic features are sent back into the ASR model and the target synthesizer sequentially to generate cyclicly reconstructed acoustic features. The reconstruction loss is adopted to optimize the target synthesizer. By recycling the pseudo-source acoustic features, this method optimizes the conversion from pseudo-source speakers to the target speaker at the training stage, which avoids the inconsistency issue. Speaker adversarial learning is also utilized for model pre-training in our implementation. Different from CycleGAN-VC [20] and StarGAN-VC [21] which are designed for one-to-one or many-to-many VC and can not be extended to unseen speakers, this method leverages the ASR model as an explicit and robust linguistic content extractor and can be directly applied to the task of any-to-one VC. Experimental results demonstrated the superiority of cyclic training on achieving consistent performance of objective evaluation metrics for various source speakers. Our

This work was supported in part by the National Nature Science Foundation of China under Grant 61871358.

proposed method also achieved better similarity of the converted speech than the baseline Rec-Syn method, especially when the pairs of source and target speakers have distant speaker characteristics.

2. METHODOLOGY

2.1. Baseline method

The baseline Rec-Syn VC method [19] is composed of four components, an ASR model, a synthesizer, a speaker encoder and a neural vocoder. The ASR model $\boldsymbol{R}$ extracts content-related features $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_t]$ from acoustic features $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_t]$ as

$$\boldsymbol{H} = \boldsymbol{R}(\boldsymbol{A}), \tag{1}$$

where t denotes the number of frames in an utterance. The speaker encoder $\boldsymbol{E}_s$ produces an utterance-level speaker code vector that provides speaker information as

$$\boldsymbol{e} = \boldsymbol{E}_s(\boldsymbol{A}). \tag{2}$$

The synthesizer $\boldsymbol{S}$ is built with an encoder-decoder architecture, which transforms the concatenated content-related features and the speaker code into acoustic features as

$$\widehat{\boldsymbol{A}} = \boldsymbol{S}(\boldsymbol{H}, \boldsymbol{e}), \tag{3}$$

where $\boldsymbol{e}$ is L2 normalized before input into $\boldsymbol{S}$.

A two-stage training strategy including pre-training and fine-tuning is adopted. The pre-training utilizes a multi-speaker training corpus where each speaker is treated as a target speaker and each utterance is assigned a speaker identity label. The parameters of the speaker encoder and the synthesizer are jointly optimized with a reconstruction loss L_{RC} as

$$L_{RC} = \frac{1}{t}\sum_{i=1}^{t} \| \widehat{\boldsymbol{a}}_i - \boldsymbol{a}_i \|_1 , \tag{4}$$

where $\widehat{\boldsymbol{a}}_i$ and $\boldsymbol{a}_i$ are the i-th frame of $\widehat{\boldsymbol{A}}$ and $\boldsymbol{A}$ respectively. When calculating $\widehat{\boldsymbol{A}}$ following Eq. (3), both the content-related features and the speaker code are from the acoustic features of each utterance in the multi-speaker corpus. Meanwhile, the speaker encoder is optimized with a speaker classification loss L_{SC} as

$$L_{SC} = \boldsymbol{CE}(\boldsymbol{p}, softmax(\boldsymbol{V}\boldsymbol{e})), \tag{5}$$

where $\boldsymbol{p}$ is the ground-truth speaker identity label in an one-hot code fashion, $\boldsymbol{V}$ is a trainable weight matrix and $\boldsymbol{CE}(.)$ represents the cross entropy function. The fine-tuning process only utilizes the data of the target speaker to achieve any-to-one VC. The speaker code of the target speaker is generated by averaging the outputs of the pre-trained speaker encoder for all training utterances of the target speaker and is not updated during fine-tuning.

At the conversion stage, the content-related features extracted from any source speech are concatenated with the speaker code of the target speaker and are sent into the synthesizer. In this paper, a Parallel WaveGAN [26] neural vocoder is used to reconstruct waveforms from the converted acoustic features.

2.2. Cyclic training of Rec-Syn VC

The cyclic training method is proposed to train the synthesizer of the target speaker in any-to-one VC. Its flowchart is shown in Figure 1. In addition to the training data of the target speaker, a multi-speaker reference corpus is employed. In every training step, for each utterance of the target speaker, a reference utterance is randomly sampled from the reference corpus. Let $\boldsymbol{A}$ denote the acoustic features of the target utterance and $\boldsymbol{A}^{ref}$ denote the acoustic features of the sampled reference utterance. Then, we will explain how to derive the cyclic reconstruction loss using $\boldsymbol{A}$ and $\boldsymbol{A}^{ref}$.

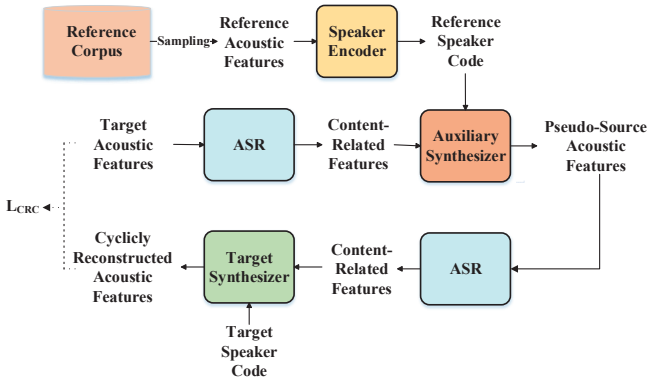

Fig. 1. The flowchart of cyclic training.

First, pseudo-source acoustic features $\boldsymbol{A}^{pse}$ which has identical linguistic contents and temporal structure with $\boldsymbol{A}$ and similar speaker characteristics with $\boldsymbol{A}^{ref}$ are generated by leveraging the ASR model $\boldsymbol{R}$, the pre-trained speaker encoder $\boldsymbol{E}_s$ and an auxiliary synthesizer $\boldsymbol{S}_{aux}$ as

$$\boldsymbol{e}^{ref} = \boldsymbol{E}_s(\boldsymbol{A}^{ref}), \tag{6}$$

$$\boldsymbol{A}^{pse} = \boldsymbol{S}_{aux}(\boldsymbol{H}, \boldsymbol{e}^{ref}), \tag{7}$$

where $\boldsymbol{H}$ denotes the content-related features of $\boldsymbol{A}$ calculated as Eq. (1). $\boldsymbol{S}_{aux}$ is utilized to generate pseudo-source acoustic features, and the generated speaker characteristics are diverse and different from the target speaker, by sampling diverse reference utterances. It won't be a problem that the generated speaker characteristics are not definitely similar to the reference speakers. In our implementation, we just employ the pre-trained synthesizer in the baseline method as $\boldsymbol{S}_{aux}$, since it has already shown good conversion quality.

Then, $\boldsymbol{A}^{pse}$ is sent back into $\boldsymbol{R}$ and the target synthesizer $\boldsymbol{S}_{trg}$ to generate cyclicly reconstructed acoustic features $\boldsymbol{A}^{cyc}$ as

$$\boldsymbol{H}^{pse} = \boldsymbol{R}(\boldsymbol{A}^{pse}), \tag{8}$$

$$\boldsymbol{A}^{cyc} = \boldsymbol{S}_{trg}(\boldsymbol{H}^{pse}, \boldsymbol{e}_{trg}), \tag{9}$$

where $\boldsymbol{e}_{trg}$ denotes the speaker code of the target speaker, which is generated by averaging the outputs of the pre-trained speaker encoder for all target training utterances.

Finally, the cyclic reconstruction loss is defined as

$$L_{CRC} = \frac{1}{t}\sum_{i=1}^{t} \| \boldsymbol{a}_i^{cyc} - \boldsymbol{a}_i \|_1 , \tag{10}$$

where $\boldsymbol{a}_i^{cyc}$ and $\boldsymbol{a}_i$ are the i-th frame of $\boldsymbol{A}^{cyc}$ and $\boldsymbol{A}$ respectively. This loss only optimizes the parameters of $\boldsymbol{S}_{trg}$ and the parameters of other models are fixed during fine-tuning.

In general, our cyclic reconstruction process is composed of two conversion flows, i.e., target to pseudo-source, and pseudo-source to target. By directly optimizing the target synthesizer with the conversion flow of pseudo-source to target and sampling reference utterances with different speaker characteristics, the inconsistency issue of conventional Rec-Syn VC can be resolved.

2.3. Speaker adversarial learning for pre-training

Speaker adversarial learning is also employed to pre-train the target synthesizer with the multi-speaker corpus before applying the proposed cyclic training method. Similar to previous work [24], an adversarial speaker classifier takes the hidden representations $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_M]$ calculated by the encoder of $\boldsymbol{S}_{trg}$ as input where M denotes the frame number of $\boldsymbol{Z}$, and produces the probability distributions $\widehat{\boldsymbol{P}} = [\widehat{\boldsymbol{p}}_1, \widehat{\boldsymbol{p}}_2, \ldots, \widehat{\boldsymbol{p}}_M]$ of speaker identities at each frame. The speaker classifier is trained with a speaker classification loss as

$$L_{CLA} = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{CE}(\widehat{\boldsymbol{p}}_i, \boldsymbol{p}), \tag{11}$$

where $\boldsymbol{p}$ denotes the ground-truth one-hot speaker identity label. The synthesizer is trained with the reconstruction loss and the adversarial loss as

$$L = L_{RC} - \lambda L_{CLA}, \tag{12}$$

where λ is the weight of the adversarial loss. The speaker classifier and synthesizer are alternately updated. The adversarial loss of maximizing L_{CLA} is expected to fool the classifier, resulting in speaker-independent encoder outputs in the pre-trained synthesizer.

3. EXPERIMENTS

3.1. Experimental conditions

We evaluated the performance of our proposed method on an internal Chinese multi-speaker dataset, which included 45 speakers with non-parallel recordings. One female and one male speaker (denoted as *TF* and *TM*) were set as the target speaker of any-to-one VC respectively, with 300 utterances for training, 50 for validation and 50 for test. 24 speakers were randomly selected from the other 43 speakers for model pre-training, which contained 140-hours speech data. For each speaker, the evaluation set contained 50 utterances, and the remaining was used for training. The remaining 19 speakers unseen in the entire training process were set as source speakers to evaluate the performance of any-to-one VC.

3.2. Implementations

All models adopted the same 80-dimensional mel-spectrograms extracted every 10 ms as acoustic features. The joint CTC-attention based approach similar to previous work [27] was employed to build the ASR model that was composed of 2 VGG-like CNN layers with 512 units, 4 BiLSTM layers with 256 units per direction, location-sensitive attention [28] and one-layer LSTM-based decoder with 512 units. 2200 hours of Chinese recordings were used for training. 512-dimensional bottleneck features, i.e, the hidden representation of the last BiLSTM layer, were taken as content-related features. The structures of the speaker encoder and the synthesizer followed our previous work [19]. The speaker adversarial classifier was built with 3 CNN layers and a fully-connected layer.

The synthesizer were optimized with Adam [29] optimizer with batch size 40 for pre-training and 10 for fine-tuning. When pre-training, the learning rate was set as 0.001 for the first 70K iterations and then exponentially decayed by 0.7 for every 5K iterations. The pre-training finished after 100K iterations. The weight λ in Eq. (12) was set following previous work [24]. During fine-tuning, the multi-speaker data used for pre-training was also utilized as the reference corpus in Figure 1. It should be noted that the proposed and the baseline methods for comparison shared the same model structures as above.

Table 1. MCDs (dB) and F_0 RMSEs (Hz) of any-to-one VC from 19 test source speakers using baseline and proposed methods. The means and standard deviations are calculated among the average MCDs and F0 RMSEs of each source speaker. "baseline+adv" denotes the baseline method with speaker adversarial learning for pre-training.

Female Target Speaker (*TF*)				
Method	Mean		Standard Deviation	
	MCD	F_0 RMSE	MCD	F_0 RMSE
baseline	**3.215**	25.731	0.079	5.679
baseline+adv	3.313	24.478	0.071	3.338
proposed	3.289	**20.103**	**0.020**	**1.386**
Male Target Speaker (*TM*)				
Method	Mean		Standard Deviation	
	MCD	F_0 RMSE	MCD	F_0 RMSE
baseline	4.003	36.542	0.083	6.004
baseline+adv	4.058	35.493	0.061	4.453
proposed	**3.999**	**30.564**	**0.022**	**1.445**

We followed an open-source implementation of Parallel WaveGAN [1] to train the vocoder using the 300 training utterances of the target speaker [2].

3.3. Objective evaluation

Due to the lack of parallel test data, we generated pseudo-source acoustic features parallel to the test utterances of the target speaker for objective evaluation. For each of the 19 test source speakers, we treated it as a conversion target and fine-tuned a synthesizer using its 300 utterances after pre-training with speaker adversarial learning. Then, the test utterances of the target speaker are converted towards each of these source speakers to produce pseudo-source acoustic features, which were used as the input to generate the converted speech of the target speaker for evaluation.

STRAIGHT [30] was adopted to extract mel-cepstral coefficients (MCCs) and F_0 from the natural and converted speeches. Mel-cepstrum distortion (MCD) and root mean square error of F_0 (F_0 RMSE) were used as objective metrics. In addition to the baseline method introduced in Section 2.1, the method (named "*baseline+adv*") which applied the speaker adversarial pre-training in Section 2.3 to the baseline was also included for comparison. The means and standard deviations of MCD and F_0 RMSE among 19 test source speakers were calculated and shown in Table 1.

We can see from Table 1 that the speaker adversarial pre-training helped the baseline model to achieve more accurate F_0 prediction with better consistency across source speakers, and our proposed method achieved lower mean F_0 RMSE than both baseline methods for both target speakers. The proposed method achieved slightly lower mean MCD than both baseline methods for both target speakers except the baseline method without speaker adversarial learning for the target speaker *TF*. Furthermore, our proposed method achieved much lower standard deviations of MCD and F_0 RMSE than the two baseline methods for both target speakers, which demonstrated its effectiveness on achieving consistent performance of acoustic feature prediction for different source speakers.

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

[2] Audio samples are available at https://nian2932491631.github.io/CycleRecSynVC/.

Table 2. Mean opinion scores (MOS) and their 95% confidence intervals of baseline and proposed methods. The numbers brackets indicate the cosine similarity between the speaker codes of source and target speakers.

Female Target Speaker (*TF*)			
Source	Method	Naturalness	Similarity
SF1(0.114)	*baseline*	3.975±0.140	4.075±0.115
	proposed	3.917±0.130	**4.200**±0.116
SF2(0.098)	*baseline*	3.842±0.128	3.608±0.137
	proposed	3.800±0.142	**3.950**±0.144
SM1(0.076)	*baseline*	3.817±0.137	3.967±0.140
	proposed	3.933±0.144	**4.133**±0.124
SM2(-0.127)	*baseline*	3.725±0.141	3.058±0.131
	baseline+adv	3.833±0.133	3.575±0.116
	proposed	3.842±0.150	**3.908**±0.122
	proposed-adv	3.892±0.128	3.717±0.105
Male Target Speaker (*TM*)			
Source	Method	Naturalness	Similarity
SF1(-0.102)	*baseline*	3.208±0.119	2.442±0.107
	baseline+adv	3.492±0.122	3.075±0.108
	proposed	3.458±0.122	**3.525**±0.094
	proposed-adv	**3.567**±0.126	3.250±0.115
SF2(-0.049)	*baseline*	3.425±0.146	3.025±0.166
	proposed	**3.683**±0.128	**3.533**±0.102
SM1(0.084)	*baseline*	3.600±0.149	3.167±0.137
	proposed	3.583±0.140	**3.508**±0.112
SM2(0.143)	*baseline*	3.750±0.127	3.458±0.093
	proposed	3.700±0.149	**3.767**±0.134

3.4. Subjective evaluation

We selected four speakers (two females and two males, denoted as *SF1*, *SF2*, *SM1* and *SM2*) from the 19 test source speakers and used their natural recordings as source speeches for subjective evaluation. Since the speaker codes are all L2 normalized, we calculated the cosine similarity between the speaker codes of source and target speakers to measure the distance of their speaker charcterisitcs , and the results are shown in the first column of Table 2. We can see that the cosine similarity of *SM2-TF* pair and that of *SF1-TM* pair were the lowest ones for the two target speakers, which implied that these two pairs of source and target speakers had the most dissimilar speaker characteristics. We confirmed this by listening to their natural recordings.

For each of the eight source-target pairs, the proposed method and the baseline method are grouped into a mean opinion score (MOS) listening test to compare their naturalness and similarity performance. Especially for *SM2-TF* and *SF1-TM* pairs, the proposed method was compared with not only the baseline method, but also *baseline+adv* and an ablated method (named "*proposed-adv*") which employed the same pre-training as the baseline method without speaker adversarial learning. In each listening test, 10 utterances were randomly selected from the test set and 12 Chinese listeners were involved. Every listener was asked to give a 5-scale opinion score (5: excellent, 4: good, 3: fair, 2: poor, 1: bad) on both naturalness and similarity of each sample. The results are shown in Table 2.

Our proposed method achieved significantly higher naturalness for *SF1-TM* and *SF2-TM* pairs than the baseline method with p-values of paired t-test 6.5×10^{-3} and 4.7×10^{-3} respectively.

Table 3. MCDs (dB) and F_0 RMSEs (Hz) of *SM2-TF* and *SF1-TM* pairs using baseline and proposed methods.

Method	*SM2-TF*		*SF1-TM*	
	MCD	F_0 RMSE	MCD	F_0 RMSE
baseline	3.293	46.817	4.259	55.315
baseline+adv	3.354	37.211	4.266	48.812
proposed	3.290	**25.309**	4.103	**38.813**
proposed-adv	**3.139**	26.699	**4.085**	40.982

For the other six speaker pairs, there were no significant differences on naturalness ($p > 0.05$) between our proposed method and the baseline method. Regarding with similarity, our proposed method outperformed the baseline method for all speaker pairs significantly ($p < 0.05$).

For the *SM2-TF* and *SF1-TM* pairs where the source speaker had the most dissimilar voice to the target speaker, our proposed method improved the similarity MOS of the baseline method with a large margin. This improvement was much larger than the ones achieved for other pairs. According to the results in Table 2, the similarity of our proposed method was also significantly higher than *baseline+adv* for these two pairs. Since both the proposed method and *baseline+adv* employed the same pre-training strategy, this result indicates the superiority of our proposed cyclic training method. Comparing the ablated model *proposed-adv* with *baseline* and *baseline+adv*, we can see that the ablated model achieved the highest similarity MOS. This means that the proposed cyclic training method was more effective than speaker adversarial pre-training on improving the similarity of these two pairs. However, the performance of the ablated model was still not as good as the proposed one. This demonstrated the feasibility and advantage of combining cyclic training with speaker adversarial pre-training. For further analysis, the objective evaluation results of these two pairs are compared among different methods as shown in Table 3. We can see that *proposed* and *proposed-adv* achieved slightly lower MCD and much lower F_0 RMSE than *baseline* and *baseline+adv*. Comparing *proposed-adv* with *baseline* and comparing *proposed* with *baseline+adv*, we can see that cyclic training improved the accuracy of acoustic feature prediction for these two pairs no matter whether the speaker adversarial pre-training was applied. These results are consistent with the subjective similarity results.

4. CONCLUSION

In this paper, a novel cyclic training method is proposed to address the inconsistency between the content-related features used at the training stage and the test stage for recognition-synthesis based any-to-one voice conversion. A multi-speaker reference corpus is adopted to generate pseudo-source acoustic features with diverse speaker characteristics and parallel linguistic contents with the training data of the target speaker, which are further used to extract content-related features to fine-tune the target synthesizer in a cyclic way. Experimental results demonstrated the superiority of our proposed method on achieving consistent accuracy of acoustic feature prediction for various source speakers and on improving the similarity of the converted speech. To integrate prosody transfer into cyclic training for controllable any-to-one voice conversion will be a task of our future work.

5. REFERENCES

[1] D. G. Childers, K. Wu, D. M. Hicks, and B. Yegnanarayana, "Voice conversion," *Speech Communication*, vol. 8, no. 2, pp. 147–158, 1989.

[2] D. G. Childers, B. Yegnanarayana, and K. Wu, "Voice conversion: Factors responsible for quality," in *ICASSP*, 1985, pp. 748–751.

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

[4] Z. Wu, E. S. Chng, and H. Li, "Exemplar-based voice conversion using joint nonnegative matrix factorization," *Multimedia Tools and Applications*, vol. 74, no. 22, pp. 9943–9958, 2015.

[5] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2009.

[6] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *ICASSP*, 2009, pp. 3893–3896.

[7] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, no. 5, pp. 954–964, 2010.

[8] L.-H. Chen, Z.-H. Ling, Y. Song, and L.-R. Dai, "Joint spectral distribution modeling using restricted boltzmann machines for voice conversion." in *INTERSPEECH*, 2013, pp. 3052–3056.

[9] Z. Wu, E. S. Chng, and H. Li, "Conditional restricted boltzmann machine for voice conversion," in *IEEE China Summit and International Conference on Signal and Information Processing*, 2013, pp. 104–108.

[10] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion using speaker-dependent conditional restricted boltzmann machine," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–12, 2015.

[11] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *ICASSP*, 2015, pp. 4869–4873.

[12] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 631–644, 2019.

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

[14] S. Liu, J. Zhong, L. Sun, X. Wu, X. Liu, and H. Meng, "Voice conversion across arbitrary speakers based on a single target-speaker utterance," in *INTERSPEECH*, 2018, pp. 496–500.

[15] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "WaveNet vocoder with limited training data for voice conversion." in *INTERSPEECH*, 2018, pp. 1983–1987.

[16] S. H. Mohammadi and T. Kim, "One-shot voice conversion with disentangled representations by leveraging phonetic posteriorgrams," in *INTERSPEECH*, 2019, pp. 704–708.

[17] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *ICASSP*, 2019, pp. 6790–6794.

[18] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *ICASSP*, 2018, pp. 5274–5278.

[19] L.-J. Liu, Y.-N. Chen, J.-X. Zhang, Y. Jiang, Z.-H. Ling, and L.-R. Dai, "Non-parallel voice conversion with autoregressive conversion model and duration adjustment," in *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020.

[20] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2100–2104.

[21] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 266–273.

[22] 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 *APSIPA*, 2016, pp. 1–6.

[23] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019.

[24] S. Ding, G. Zhao, and G.-O. Ricardo, "Improving the speaker identity of non-parallel many-to-many voice conversion with adversarial speaker recognition," in *INTERSPEECH*, 2020, pp. 776–780.

[25] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Recognition-synthesis based non-parallel voice conversion with adversarial learning," in *INTERSPEECH*, 2020, pp. 771–775.

[26] 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 *ICASSP*, 2020, pp. 6199–6203.

[27] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *ICASSP*, 2017, pp. 4835–4839.

[28] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

[30] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction," *Speech Communication*, vol. 27, no. 34, pp. 187–207, 1999.