Odyssey 2020 The Speaker and Language Recognition Workshop
1-5 November 2020, Tokyo, Japan

Generative Adversarial Networks for Singing Voice Conversion with and without Parallel Data

Berrak Sisman [1,2] , *Haizhou Li* [2]

[1] Information Systems Technology and Design Pillar, Singapore University of Technology and Design
[2] Dept. of Electrical and Computer Engineering, National University of Singapore, Singapore

berraksisman@u.nus.edu, haizhou.li@nus.edu.sg

Abstract

Singing voice conversion (SVC) is a task to convert one singer's voice to sound like that of another, without changing the lyrical content. Singing conveys lexical and emotional information through words and tones. Both are to be transferred from the source to target. In this paper, we propose novel solutions to SVC based on Generative Adversarial Networks (GANs) with and without parallel training data. With parallel data, we employ GANs to minimize the differences of the distributions between the original target parameters and the generated singing parameters. With non-parallel training data, we employ CycleGANs to estimate an optimal pseudo pair between source and target singers. Moreover, the proposed solutions perform well with limited amount of training data. The experiments show that (1) GANs outperform other state-of-the-art voice conversion when parallel training data are available, (2) CycleGANs achieve competitive voice conversion quality without the need of parallel training data. To our best knowledge, this paper is the first to study the use generative adversarial networks for singing voice conversion with and without parallel data.

1. Introduction

It is not an easy task for one to mimic another's singing. While professional singers are trained to control and vary their vocal timbres, they are bounded by the physical limit of human vocal production system. As illustrated in Figure 1, singing voice conversion provides an extension to one's vocal ability to control the voice beyond the physical limit and express in an extended variety of ways. Singing voice conversion (SVC) has many practical applications such as singing synthesis, dubbing of sound track, and perfecting one's singing. We note that singing voice conversion is challenging because singing is a form of art and any distortion on the converted singing voice cannot be tolerated.

Singing voice conversion is a task that is technically related to singing voice synthesis and speech voice conversion. Recently, singing voice synthesis [1–5] technology has advanced tremendously, from unit selection, singing style modeling, to novel vocoding processing, that our research can benefit from. Singing voice conversion also shares similar motivation with the conventional speech voice conversion [6–13] where we transform the person-dependent traits from source to target and carry over the person-independent content.

However, SVC differs from speech voice conversion in many ways. For example, in speech voice conversion, speech prosody, that includes pitch, dynamics, duration of words, etc., describes speaker individuality. Therefore, it should be transformed from the source to the target speaker [14–17]. However, in SVC, the manner of singing is primarily determined by the sheet music itself, therefore, is considered as person-independent. As a result, only the characteristics of voice identity, such as the timbre, are considered as the person-dependent traits to be converted [18–22]. Hence, we focus on spectrum conversion in this paper.

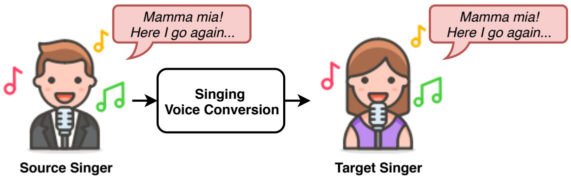

Figure 1: Singing voice conversion is to convert one's singing to that of another. This paper studies the use of generative adversarial network with and without parallel training data.

With parallel training data of two singers, the spectral mapping between source and target is more straightforward. Successful implementations include GMM-based direct waveform modification [18, 19, 23] technique, concatenative singing voice conversion [20], GMM-based spectrum mapping for VOCALOID singing synthesizer [24] and GMM-based voice conversion with vocal tract area function [21]. In this paper, we would like to study how we approach this problem by using GAN, a deep learning technique.

In practice, it is more convenient to train a system without the need of parallel training data [25]. Motivated by this, we also propose to train a cycle-consistency generative adversarial network for singing voice conversion on non-parallel data. With CycleGAN, the network can learn a complex relationship between source and target singing features through an adversarial process without the need of parallel data. In this way, we don't require any external process, such as alignment or Automatic Speech Recognition (ASR), to produce high-quality singing even with limited data.

Similar to the prior work [23], this paper mainly focuses on spectrum conversion as it is not always necessary to transform F0 values of the source singer to those of the target singer, because both source and target speakers often sing on key. Moreover, the conversion of aperiodicity usually only has a small impact on the converted singing voice.

The main contributions of this paper include: 1) we propose a novel singing voice conversion frameworks that are based on Generative Adversarial Networks; 2) we achieve high quality converted singing voice without any external process, such as speech recognition; 3) by using CycleGANs, we

Speech Samples: https://sites.google.com/view/berraksisman/

10.21437/Odyssey.2020-34

achieve high quality parallel-data-free singing voice that outperforms the baseline; and 4) we reduce the reliance on large amount of data for both parallel and nonparallel training scenarios. To our best knowledge, this paper reports the first successful attempt to use generative adversarial networks for parallel-data free singing voice conversion.

This paper is organized as follows: In Section 2, we explain the generative adversarial networks (GANs). In Section 3, we explain the proposed GAN-based SVC framework with parallel training data. In Section 4, we present CycleGAN-based SVC framework with non-parallel data. We report the experiments in Section 5 and conclude in Section 6.

2. Generative Adversarial Networks (GANs)

The traditional generative adversarial network [26] consists of a generator and a discriminator. In this framework, the generator learns a mapping function from the distribution of source to the distribution of target. Generative adversarial networks have recently been shown to be an effective training method and have become popular in many fields such as image generation [27], image synthesis [27], speech enhancement [28], language identification [29], and text-to-speech synthesis [30].

In speech voice conversion, GANs have been shown to be effective [31, 32] for the cases with parallel training data and achieve remarkable performance in terms of voice quality and speaker similarity. We note that we cannot always assume that parallel training data is available in practice. To be free from such assumption, some deep learning conversion frameworks such as joint usage of DBLSTM and i-vector [33], variational auto-encoder [34], DBLSTM based Recurrent Neural Networks [12, 35, 36] have been proposed. More recently, GAN-based speech voice conversion techniques such as VAW-GAN [37], and CycleGAN [38] achieve remarkable performance with nonparallel training data.

Just like in speech voice conversion, deep learning has also became popular in singing voice conversion. Generative adversarial network approach has been used for singing voice conversion under the assumption of parallel training data [39]. With non-parallel data, a VAE-based technique was proposed, that aims to disentangle singer and vocal information [40]. Different from the previous studies, we propose a generative adversarial network, that works for high-quality singing voice conversion, both with and without parallel training data.

3. Singing Voice Conversion with Parallel Data

The statistical methods such as GMM [18, 19] represent an early success of singing voice conversion. The furtherance in deep learning has a positive impact in many fields, with no exception to singing voice conversion. We propose to use GANs to learn the essential differences between the source singing and the original target singing through a discriminative process, that was previously formulated in [39] as SINGAN. In this paper, we further the study of a SINGAN as part of the GAN solutions to singing voice conversion in a comparative study.

We next explain the technical steps of SINGAN when parallel training data are available. We use source and target singing spectral features as the input for the generative adversarial network. The GAN structure consists of two DNNs [41], that are iteratively updated by minibatch stochastic gradient descent. The generator performs the spectral mapping, while the discriminator acts as a DNN-based anti-spoofing system that distinguishes between natural and synthetic singing voice.

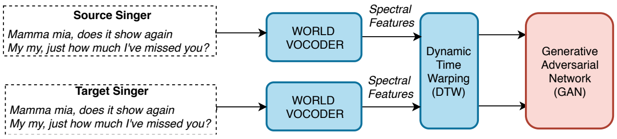
Source Singer
Mamma mia, does it show again
My my, just how much I've missed you?
Target Singer
Mamma mia, does it show again
My my, just how much I've missed you?
WORLD VOCODER
WORLD VOCODER
Spectral Features
Spectral Features
Dynamic Time Warping (DTW)
Generative Adversarial Network (GAN)

Figure 2: The training phase of the proposed GAN-based singing voice conversion framework, when source-target parallel singing data are available.

3.1. Training Process

The training phase of SINGAN is given in Figure 2. The training process involves three steps:

1. To perform WORLD analysis to obtain the spectral and prosody features,
2. To use dynamic time warping algorithm for temporal alignment of source and target singing spectral features, and
3. To train the generative adversarial network by using the aligned singing source and target features.

Previous studies [23] suggest that, in intra-gender SVC, such as male-to-male and female-to-female singer identity conversions, it is not always necessary to transform F0 values of the source singer to those of the target singer, because both singers often sing on key. Moreover, the conversion of aperiodicity usually has only a small impact on the converted singing voice. Therefore, it suffices to only perform spectral feature conversion to achieve acceptable singing voice quality.

3.2. Run-time Conversion Process

At run-time conversion, we perform three steps as follows: 1) to obtain source singing features using WORLD analysis, 2) to generate the converted singing spectral features by using the GAN, that is already trained, and 3) to generate the converted singing waveform by using WORLD synthesis.

In this paper, as stated by the previous studies, we do not perform F0 conversion for intra-gender SVC experiments [23]. However, for inter-gender SVC experiments, we perform linear F0 conversion that is to normalize the mean and variance of the source speaker's F0 to that of target speaker. In all experiments, we copy the aperiodicity of the source singing to the target.

3.3. Comparison with Related Work

SINGAN shares similar motivation with [25] regarding the use of deep learning. However, it differs from [25] in many ways, for example: 1) SINGAN adopts GAN instead of DBLSTM [25] for spectral mapping; 2) SINGAN performs one-to-one singing voice conversion with a small parallel training set, while [25] studies a many-to-one conversion with a large amount of non-parallel training data. 3) SINGAN does not depend on automatic speech recognition (ASR) performance, while [25] performs mapping in between spectral features and posteriograms, that are the intermediate results of an ASR system.

4. Singing Voice Conversion with Non-parallel Data

Cycle Consistent Adversarial Networks have shown to be effective in many applications, such as image manipulation [42–44], image synthesis [45–47], and speech voice conversion [38, 48]. We note that image-to-image translation and speech voice conversion are similar in the sense that they both learn the mapping between source domain and target domain. CycleGAN can learn from non-parallel training data, it provides a solution to model the singer translation. To our best knowledge, CycleGAN has not been studied for singing voice conversion. We next formulate the CycleGAN solution to learn the forward and inverse mappings simultaneously using adversarial and cycle-consistency losses,

CycleGAN model is trained to find an optimal pseudo pair from the unpaired singing data of two singers. The adversarial loss contributes to reducing over-smoothing of the converted feature sequence. We configure the CycleGAN with gated CNNs and train it with an identity-mapping loss. This allows us to preserve the lyrical content of the source singer. The training phase of the proposed CycleGAN is given in Figure 3. Hereafter, the source singer is denoted as *s*, and the target as *t*. Our goal is to learn a mapping from source $x_s \in X_s$ to target $y_t \in Y_t$ with non-parallel training data. We solve this problem by optimizing the three loss functions, that we discuss in the following subsections.

The proposed CycleGAN preserves the lyrical content while converting the speaker identity through a source-target mapping. It has the following advantages: 1) no need of any external processes, such as ASR, and 2) no need for any alignment technique which is very challenging for singing data and has been explored deeply by the community [49, 50].

4.1. Adversarial Loss

In singing voice conversion, our aim is to optimize the distribution of the converted singing features as much as closer to the distribution of target singer. To achieve this, we use the following objective function:

$$\begin{aligned}\mathscr{L}_{adv}(G_{X_s\to Y_t}, D_{Y_t}) &= E_{y_t\sim P_{Data(y_t)}}\left[\log D_{Y_t(y_t)}\right]\\ &+ E_{x_s\sim P_{Data(x_s)}}\left[\log\left(1 - D_{Y_t}\left(G_{X_s\to Y_t(x_s)}\right)\right)\right] \qquad (1)\end{aligned}$$

The closer the distribution of converted data becomes to that of target singer, the smaller the loss (Eq. (1)) becomes. Hence, we can achieve high speaker similarity in singing voice conversion.

4.2. Cycle-Consistency Loss

The adversarial loss only tells us whether $G_{X_s\to Y_t}$ follows the target singer's data distribution and does not help preserve the source singer's contextual information. In speech voice conversion, CycleGAN [38, 48] was designed with two additional terms, the adversarial loss $\mathscr{L}_{adv}(G_{Y_t\to X_s}, D_{X_s})$ for inverse mapping $G_{Y_t\to X_s}$, and the cycle-consistency loss. For singing voice conversion, we apply the cycle-consistency loss as follows:

$$\begin{aligned}&\mathscr{L}_{cyc}\left(G_{X_s\to Y_t}, G_{Y_t\to X_s}\right)\\ &= \mathbb{E}_{x_s\sim P_{Data}(x_s)}\left[||G_{Y_t\to X_s}\left(G_{X_s\to Y_t(x_s)}\right) - x_s||_1\right]\\ &+ \mathbb{E}_{y_t\sim P_{Data}(y_t)}\left[||G_{X_s\to Y_t}\left(G_{Y_t\to X_s(y_t)}\right) - y_t||_1\right] \qquad (2)\end{aligned}$$

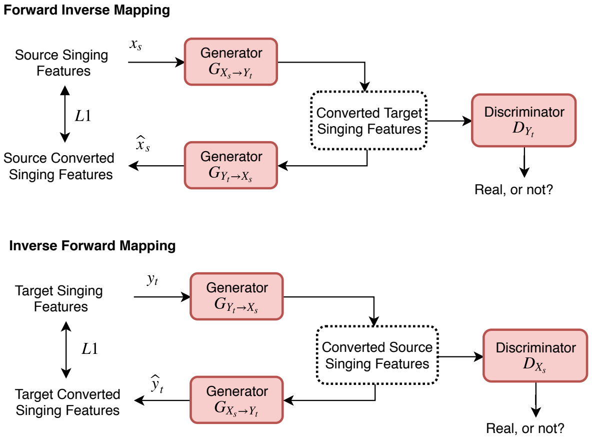
Forward Inverse Mapping
Source Singing Features
x_s
Generator $G_{X_s\to Y_t}$
Converted Target Singing Features
Discriminator D_{Y_t}
Real, or not?
$L1$
$\hat{x}_s$
Source Converted Singing Features
Generator $G_{Y_t\to X_s}$
Inverse Forward Mapping
Target Singing Features
y_t
Generator $G_{Y_t\to X_s}$
Converted Source Singing Features
Discriminator D_{X_s}
Real, or not?
$L1$
$\hat{y}_t$
Target Converted Singing Features
Generator $G_{X_s\to Y_t}$

Figure 3: Training phase of CycleGAN with cycle-consistency loss with non-parallel training data for singing voice conversion.

These additional terms $G_{X_s\to Y_t}$ and $G_{Y_t\to X_s}$ helps maintain the contextual information between the source and target pair (x_s, y_t).

4.3. Identity-Mapping Loss

A cycle-consistency loss provides constraints on a structure; however, it would not suffice to guarantee that the mappings always preserve lyrical content of the source singer. To explicitly preserve the lyrical content, we incorporate an identity-mapping loss.

$$\begin{aligned}&\mathscr{L}_{id}\left(G_{X_s\to Y_t}, G_{Y_t\to X_s}\right)\\ &= \mathbb{E}_{x_s\sim P_{Data}(x_s)}\left[||G_{Y_t\to X_s}(x_s) - x_s||\right]\\ &+ \mathbb{E}_{y_t\sim P_{Data}(y_t)}\left[||G_{X_s\to Y_t}(y_t) - y_t||\right] \qquad (3)\end{aligned}$$

The studies on CycleGAN have showed the effectiveness of this loss for color preservation in image-to-image translation and linguistic content preservation in mono-lingual [38] and cross-lingual voice conversion [51]. Hence, we have good reason to expect that it preserves the lyrical content from the source speaker in singing voice conversion.

5. Experiments

We perform objective and subjective evaluations on NUS Sung and Spoken Lyrics Corpus (NUS-48E corpus) [52]. This corpus consists of audio recordings of the sung and spoken lyrics of 48 English songs by 12 professional singers. For both parallel and non-parallel training data settings, we conduct experiments with 3 and 5 source-target singing pairs. In all experiments, we use the WORLD vocoder for feature analysis and synthesis. We first extract 34 Mel-cepstral coefficients (MCEPs), logarithmic fundamental frequency (log F0), and aperiodicities (APs) every 5 ms by using the WORLD analyzer. We then normalize the source and target MCEPs to zero-mean and unit variance by using the statistics of the training sets. The silent frames are removed from the training data in order to increase training accuracy. We only perform conversion on MCEPs, hence, the objective of our experiments is to analyze the quality of the converted MCEPs.

Frameworks	Gender	Training Data	MCD
DNN Baseline	m-to-m	Parallel 3-3	5.98
		Parallel 5-5	5.73
	f-to-m	Parallel 3-3	6.12
		Parallel 5-5	5.89
GAN	m-to-m	Parallel 3-3	5.52
		Parallel 5-5	5.36
	f-to-m	Parallel 3-3	5.74
		Parallel 5-5	5.53
CycleGAN	m-to-m	Non-parallel 3-3	5.94
		Non-parallel 5-5	5.72
	f-to-m	Non-parallel 3-3	6.09
		Non-parallel 5-5	5.85

Table 1: A comparison between the proposed GAN, CycleGAN, and the DNN baseline for singing voice conversion, that are trained on parallel and non-parallel training data.

5.1. Experimental Conditions

5.1.1. *DNN-based singing voice conversion*

We note that DNN-based approach [41] has been widely used as a baseline for some GAN-based speech synthesis [30] and voice conversion frameworks [39]. Therefore, we also choose to use deep neural network (DNN) approach to singing voice conversion. Our aim is to find a mapping between source and target singers by using parallel training data. During training, we first use a dynamic time warping algorithm to temporally align source and target singing features. We then train a DNN by using these aligned source and target singing features, that is in a similar way to the conventional speech voice conversion. The hidden layers of the DNN have $3 * 512$ units.

5.1.2. *Singing voice conversion with parallel data (GAN)*

The proposed GAN structure consists of two DNNs, that are iteratively updated by minibatch stochastic gradient descent. In the experiments, we construct DNNs for male-to-male (denoted as m-to-m) and female-to-male (denoted as f-to-m) singing voice conversion. The hidden layers of the generator and discriminator have $3 * 512$ units and $3 * 256$ units, respectively. The discriminator, that we use in this paper can be seen as a DNN-based anti-spoofing system that distinguishes between natural and synthetic singing voice.

5.1.3. *Singing voice conversion with non-parallel data (CycleGAN)*

In CycleGAN, we design the generator using a one-dimensional (1D) CNN [53, 54] to capture the relationship among the overall features while preserving the temporal structure. The target of the generator is also the 513-dimension spectral envelope. We design the discriminator using a 2D CNN to focus on a 2D spectral texture. As a pre-process, we normalize the source and target MCEPs per dimension. We set λ_{cyc} =10 and λ_{id} =5. We train the network using the Adam optimizer with a batch size of 1. We empirically set the initial learning rates to 0.0002 for the generator and 0.0001 for the discriminator. Consistent with DNN and GAN, we construct CycleGAN for male-to-male (denoted as m-to-m) and female-to-male (denoted as f-to-m) singing voice conversion. We note that CycleGAN doesn't require source-target pair to have the same length. Therefore, time alignment is not necessary.

DNN	CycleGAN	GAN
(3.29 ± 0.31) %	(3.31 ± 0.26) %	(3.85 ± 0.27) %

Table 2: Mean Opinion Score (MOS) comparison of DNN, CycleGAN and GAN for singing voice conversion. DNN and GAN frameworks are trained on 5-5 parallel source-target song pairs, while CycleGAN is trained on same amount of non-parallel data.

5.2. Objective Evaluation

We adopt the Mel-cepstral distortion (MCD) [23] between the MCEPs of target singer's natural singing and the converted ones. We note that a lower MCD value indicates smaller spectral distortion, hence better conversion performance.

In Table I, we report the MCD values under different training settings. We would like to compare the following frameworks: 1) CyleGAN with nonparallel training data, 2) GAN with parallel training data, and 3) DNN baseline with parallel training data. Firstly, we observe that GAN outperforms the DNN-based approach by achieving lower MCD in all cases. The results suggest that GAN framework generates a singing spectrum, that is more similar to the original target singer than the DNN-based framework. We also would like to note that the proposed GAN can work remarkably well with limited amount of parallel data. For example, GAN (male-to-male) with 3 song pairs achieves the MCD value of 5.52, while baseline DNN (male-to-male) achieves the MCD value of 5.73 even with 5 song pairs.

We further compare CycleGAN for parallel-data-free SVC with GAN and the baseline DNN. Table 1 shows that CycleGAN is as competitive as the DNN baseline, as well as GAN. We would like to note that CycleGAN is trained under a more competitive scenario, where parallel data are not available. For example, CycleGAN (female-to-male) with 3 song pairs achieves the MCD value of 6.09, while baseline DNN (female-to-male) achieves the MCD value of 6.12 and GAN achieves 5.74. The results also suggest that the proposed CycleGAN framework eliminates the reliance on parallel training data or any alignment technique, hence achieves high quality singing voice.

5.3. Subjective Evaluation

We further conduct three listening experiments to assess the performance of generative adversarial networks for singing voice conversion in terms of voice quality and speaker similarity. 15 subjects participate in the listening tests and each subject listens to 50 converted singing voices.

We first evaluate the sound quality of the converted voices with mean opinion score (MOS) between DNN, GAN and CycleGAN for singing voice conversion, that is reported in Table 2. The listeners rate the quality of the converted voice using a 5-point scale: "5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for bad. We note that DNN and GAN are trained with 5 parallel songs, while CycleGAN is trained with 5 nonparallel songs. It is observed that GAN outperforms both DNN and CycleGAN counterparts. It is worth mentioning that CycleGAN with non-parallel data achieves better results than DNN with parallel-data, that we believe remarkable.

We conduct preference test, that is reported in Figure 4, to compare CycleGAN and GAN for SVC, in terms of speaker similarity. We note that this is not a fair comparison as CyleGAN is trained with nonparallel data. It is observed that GAN

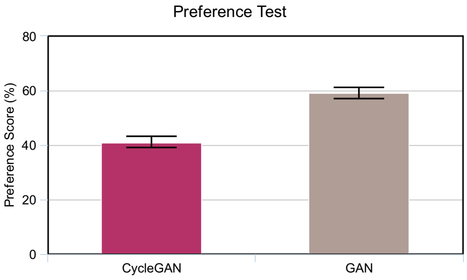

Figure 4: The preference test for speaker similarity between the proposed CycleGAN and GAN for singing voice conversion. CycleGAN is trained with 5-5 non-parallel song pairs while GAN is trained with 5-5 parallel song pairs.

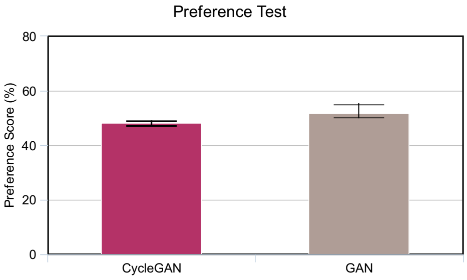

Figure 5: The preference test for speaker similarity between the proposed CycleGAN and GAN for singing voice conversion. CycleGAN is trained with 5-5 non-parallel song pairs while GAN is trained with 3-3 parallel song pairs.

outperforms CycleGAN when they are both trained with the same amount of data, that is 5 songs from source and target.

We further conduct another speaker similarity preference test, that is reported in Figure 5, to compare CycleGAN and GAN with different amounts of training data. To show the capability of CycleGAN, we use 5 nonparallel song pairs for CycleGAN training, while we only use 3 parallel song pairs for GAN training. We show that CycleGAN with nonparallel singing data achieves comparable results to GAN with parallel singing data, as it is chosen as the better sample for (48.1 ± 1.3) percent of the time. This experiment suggests that CycleGAN trained on non-parallel data can achieve similar voice conversion quality as GAN trained on parallel data.

6. Conclusion

In this paper, we propose a novel solution based on Generative Adversarial Networks (GANs) to singing voice conversion with and without parallel training data. By using GANs, we minimize the differences of the distributions between the original target singing spectrum and the converted singing spectrum. The proposed GAN framework achieves high quality converted singing voice with parallel training data. By using CycleGANs, we estimate an optimal pseudo pair between source and target singers, even from nonparallel training data. Furthermore, we also show that the proposed GAN frameworks perform better with less training data than other deep neural networks. With or without parallel training data available, generative adversarial networks outperform DNN baseline by achieving high-quality synthesized singing voice.

We have also tried applying generative adversarial networks for many-to-many singing voice conversion and have obtained some good preliminary results. More investigation will be conducted in the future.

7. Acknowledgement

This work is supported by Human-Robot Interaction Phase 1 (Grant No. 192 25 00054), National Research Foundation Singapore under the National Robotics Programme. It is also supported by National Research Foundation Singapore under the AI Singapore Programme (Award Number: AISG-100E-2018-006), and Programmatic Grant No. A18A2b0046 (Human Robot Collaborative AI for AME) and A1687b0033 (Neuromorphic Computing) from the Singapore Government's Research, Innovation and Enterprise 2020 plan in the Advanced Manufacturing and Engineering domain.

8. References

[1] K Saino, M Tachibana, and H Kenmochi, "A singing style modeling system for singing voice synthesizers," *INTERSPEECH*, 2010.

[2] M Nishimura, K Hashimoto, O Keiichiro, N Nankaku, and K Tokuda, "Singing voice synthesis based on deep neural networks," *INTERSPEECH*, 2016.

[3] H Kenmochi and H Ohshita, "VOCALOID – Commercial singing synthesizer based on sample concatenation," *INTERSPEECH*, 2007.

[4] Xavier Rodet, "Synthesis and processing of the singing voice," *in Proc. of the 1st IEEE Benelux MPCA Workshop*, 2002.

[5] T. Nakano and M. Goto, "Vocalistener2: A singing synthesis system able to mimic a user's singing in terms of voice timbre changes as well as pitch and dynamics," *ICASSP*, 2011.

[6] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *In ICASSP*, pp. 655–658, 1988.

[7] Kiyohiro Shikano, Satoshi Nakamura, and Masanobu Abe, "Speaker Adaptation and Voice Conversion by Codebook Mapping," *IEEE International Sympoisum on Circuits and Systems*, pp. 594–597, 1991.

[8] Tomoki Toda, Alan W. Black, and Keiichi 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.

[9] E. Helander, H. Silen, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 806–817, 2012.

[10] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Exemplar-based voice conversion in noisy environment," *In IEEE SLT*, pp. 313–317, 2012.

[11] Berrak Sisman, Haizhou Li, and Kay Chen Tan, "Sparse representation of phonetic features for voice conversion with and without parallel data," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 677–684.

[12] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," *In IEEE ICME*, 2016.

[13] Berrak Sisman, Mingyang Zhang, and Haizhou Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder.," in *Interspeech*, 2018, pp. 1978–1982.

[14] Berrak Sisman, Haizhou Li, and Kay Chen Tan, "Transformation of prosody in voice conversion," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 1537–1546.

[15] Berrak Sisman, Grandee Lee, Haizhou Li, and Kay Chen Tan, "On the analysis and evaluation of prosody conversion techniques," in *2017 International Conference on Asian Language Processing (IALP)*. IEEE, 2017, pp. 44–47.

[16] Berrak Sisman and Haizhou Li, "Wavelet analysis of speaker dependent and independent prosody for voice conversion.," in *Interspeech*, 2018, pp. 52–56.

[17] Berrak Sisman, Mingyang Zhang, and Haizhou Li, "Group sparse representation with wavenet vocoder adaptation for spectrum and prosody conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1085–1097, 2019.

[18] Kazuhiro Kobayashi, Tomoki Toda, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura, "Statistical singing voice conversion based on direct waveform modification with global variance," *INTERSPEECH*, 2015.

[19] Kazuhiro Kobayashi, Tomoki Toda, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura, "Statistical Singing Voice Conversion with direct Waveform modification based on the Spectrum Differential," *INTERSPEECH*, 2014.

[20] Fernando Villavicencio and Jordi Bonada, "Applying Voice Conversion To Concatenative Singing-Voice Synthesis," *INTERSPEECH*, 2010.

[21] Y Kawakami, H Banno, and Itakura F, "GMM voice conversion of singing voice using vocal tract area function," *IEICE technical report. Speech (Japanese edition)*, 2010.

[22] H Doi, T Toda, T Nakano, M Goto, and S Nakamura, "Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system," *APSIPA ASC*, 2012.

[23] Kazuhiro Kobayashi, Tomoki Toda, and Satoshi Nakamura, "Intra-gender statistical singing voice conversion with direct waveform modification using log-spectral differential," *Speech Communication*, 2018.

[24] H Kenmochi and H Ohshita, "VOCALOID – Commercial singing synthesizer based on sample concatenation," *INTERSPEECH*, 2007.

[25] Xin Chen, Wei Chu, Jinxi Guo, and Ning Xu, "Singing Voice Conversion with Non-parallel Data," *arXiv:1903.04124 [eess.AS]*, 2019.

[26] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *NIPS Proceedings*, 2014.

[27] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," *ICCV*, 2017.

[28] Zhong Meng, Jinyu Li, Yifan Gong, and Biing-Hwang (Fred) Juang, "Cycle-Consistent Speech Enhancement," *INTERSPEECH*, 2018.

[29] Peng Shen, Xugang Lu, Sheng Li, and Hisashi Kawai, "Conditional generative adversarial nets classifier for spoken language identification," *INTERSPEECH*, 2017.

[30] Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2017.

[31] Takuhiro Kaneko, Hirokazu Kameoka, Kaoru Hiramatsu, and Kunio Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," *INTERSPEECH*, 2017.

[32] Berrak Sisman, Mingyang Zhang, Sakriani Sakti, Haizhou Li, and Satoshi Nakamura, "Adaptive wavenet vocoder for residual compensation in gan-based voice conversion," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 282–289.

[33] Jie Wu, Zhizheng Wu, and Lei Xie, "On the use of I-vectors and average voice model for voice conversion without parallel data," *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[34] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," *APSIPA ASC*, 2016.

[35] Lifa Sun, Hao Wang, Shiyin Kang, Kun Li, and Helen Meng, "Personalized, cross-lingual TTS using phonetic posteriorgrams," *In INTERSPEECH*, pp. 322–326, 2016.

[36] Xiaohai Tian, Xu Haihua Wang, Junchao, , Eng Siong Chng, Senior Member, and Haizhou Li, "Average Modeling Approach to Voice Conversion with Non-Parallel Data," *Odyssey 2018 The Speaker and Language Recognition Workshop*, pp. 1–10, 2018.

[37] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice Conversion from Unaligned Corpora using Variational Autoencoding Wasserstein Generative Adversarial Networks," *arXiv:1704.00849 [cs.CL]*, 2017.

[38] Takuhiro Kaneko and Hirokazu Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv*, 2017.

[39] Berrak Sisman, Karthika Vijayan, Minghui Dong, and Haizhou Li, "SINGAN: Singing voice conversion with generative adversarial networks," *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2019*, , no. December, 2019.

[40] Yin-Jyun Luo, Chin-Chen Hsu, Kat Agres, and Dorien Herremans, "Singing voice conversion with disentangled representations of singer and vocal technique using variational autoencoders," *arXiv preprint arXiv:1912.02613*, 2019.

[41] Heiga Zen, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," *IEEE ICASSP*, 2013.

[42] Kenan Emir Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf Kassim, "Attribute Manipulation Generative Adversarial Networks for Fashion Images," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[43] Kenan Emir Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf Kassim, "Semantically Consistent Hierarchical Text to Fashion Image Synthesis with an enhanced-Attentional Generative Adversarial Network," *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshop*, 2019.

[44] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim, "Semantically consistent text to fashion image synthesis with an enhanced attentional generative adversarial network," *Pattern Recognition Letters*, 2020.

[45] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *ICCV*, 2017.

[46] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz, "Multimodal Unsupervised Image-to-Image Translation," *ECCV*, 2018.

[47] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman, "Toward Multimodal Image-to-Image Translation," *NIPS*, 2017.

[48] Fuming Fang, Junichi Yamagishi, Echizen I, and Jaime Lorenzo-Trueba, "High-Quality Nonparallel Voice Conversion Based on Cycle-Consistent Adversarial Network," *IEEE ICASSP*, 2018.

[49] Xiaoxue Gao, Berrak Sisman, Rohan Kumar Das, and Karthika Vijayan, "Nus-hlt spoken lyrics and singing (sls) corpus," in *2018 International Conference on Orange Technologies (ICOT)*. IEEE, 2018, pp. 1–6.

[50] Karthika Vijayan, Haizhou Li, and Tomoki Toda, "Speech-to-singing voice conversion: The challenges and strategies for improving vocal conversion processes," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 95–102, 2018.

[51] Berrak Sisman, Mingyang Zhang, Minghui Dong, and Haizhou Li, "On the study of generative adversarial networks for cross-lingual voice conversion," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019.

[52] Zhiyan Duan, Haotian Fang, Bo Li, Khe Chai Sim, and Ye Wang, "The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech," *APSIPA*, 2013.

[53] Kenan E Ak, Ashraf A Kassim, Joo Hwee Lim, and Jo Yew Tham, "Learning attribute representations with localization for flexible fashion search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7708–7717.

[54] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim, "Efficient multi-attribute similarity learning towards attribute-based fashion search," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1671–1679.