SINGAN: Singing Voice Conversion with Generative Adversarial Networks

Berrak Sisman†‡, Karthika Vijayan*, Minghui Dong † and Haizhou Li *‡
* National University of Singapore, Singapore
† Institute for Infocomm Research, A*STAR, Singapore
E-mail: berraksisman@u.nus.edu, karthika.vijayan@nus.edu.sg, haizhou.li@nus.edu.sg
† Institute for Infocomm Research, A*STAR, Singapore
E-mail: mhdong@i2r.a-star.edu.sg

Abstract—Singing voice conversion (SVC) is a task to convert the source singer’s voice to sound like that of the target singer, without changing the lyrical content. So far, most of the voice conversion studies mainly focus only on the speech voice conversion that is different from singing voice conversion. We note that singing conveys both lexical and emotional information through words and tones. It is one of the most expressive components in music and a means of entertainment as well as self expression. In this paper, we propose a novel singing voice conversion framework, that is based on Generative Adversarial Networks (GANs). The proposed GAN-based conversion framework, that we call SINGAN, consists of two neural networks: a discriminator to distinguish natural and converted singing voice, and a generator to deceive the discriminator. With GAN, we minimize the differences of the distributions between the original target parameters and the generated singing parameters. To our best knowledge, this is the first framework that uses generative adversarial networks for singing voice conversion. In experiments, we show that the proposed method effectively converts singing voices and outperforms the baseline approach.

Index Terms: Singing voice conversion, generative adversarial networks, singing voice

I. INTRODUCTION

Professional singers are believed to be good at controlling their voice timbre. However, they usually have a difficulty to change their voice to sound like that of another, due to the physical constraints of speech production. Singing voice conversion provides an extension to one’s vocal ability to control the voice beyond physical constraints and express in an extended variety of ways.

Singing voice conversion is defined as the task of converting a song of a source singer to sound like the voice of a target singer without changing the linguistic content. Singing voice conversion has seen many practical applications such as singing synthesis, dubbing of movies and enabling singers to sing songs with their desired voice timbre, etc. We note that the task of singing voice conversion is related to both singing voice synthesis and speech voice conversion.

The interest to singing voice synthesis [1]–[6] has been growing recently in the field of computer-based music technology. By entering notes and lyrics to a singing voice synthesis system, one can generate a synthesized singing voice with a specific singer’s voice identity. This new technology has created opportunities for new and innovative music products and services.

Singing voice conversion and conventional speech voice conversion are similar in many ways. Both techniques transform the person-dependent traits from source to target and carry over the person-independent content. We note that in speech voice conversion, prosody, that includes pitch, dynamics, duration of words, etc., contains important information about the speaker identity. Therefore, to achieve high quality voice conversion, the prosody should be transformed from the source speaker to the target speaker [7]–[10]. However, in singing voice conversion, the manner of singing is primarily determined by the sheet music itself, therefore, is considered as person-independent. In this case, only the characteristics of voice identity, such as the timbre, are considered as the person-dependent traits to be converted [11]–[15]. Hence, in this paper, we only focus on spectrum conversion.

The early studies on voice conversion marked a success by training a mapping function to convert the source speech to target speech with parallel training data, such as Vector Quantization (VQ) [16], codebook mapping [17], Gaussian Mixture Model (GMM) [18], partial least square regression [19], dynamic kernel partial least squares regression (DKPLS) [20], and non-negative matrix factorization (NMF) based voice conversion frameworks [9], [21], [22]. Benefiting from deep learning, voice conversion technology has advanced rapidly, providing high voice quality and speaker similarity [10], [23]–[26].

There have been some traditional statistical approaches for training a function that maps the singing vocal of a source singer to that of a target singer. The mapping function is trained to associate the spectral features between the source and target singer as illustrated in Figure 1. GMM-based direct waveform modification [11], [12], [28] technique, concate-
native singing voice conversion [13], GMM-based spectrum mapping for VOCALOID singing synthesizer [29], GMM-based voice conversion with vocal tract area function [14] and many-to-many eigenvoice conversion [15] are the traditional singing voice conversion frameworks, that achieve high-quality converted singing voice. The statistical approaches have benefited from the success of speech voice conversion.

We notice that deep learning approaches have not become very popular in singing voice conversion yet. Only recently, a DBLSTM-based singing voice conversion [30], that uses PPGs [22] for the mapping function, was studied. PPGs, derived from a speech recognition system, represent the posterior probability of the speech frame with respect to the phonetic classes, that are believed to be speaker independent [23], [31]. We note that DBLSTM with PPGs is a clever solution to speech and singing [30] voice conversion with non-parallel training data. However, it works with a speech recognition system, therefore, hinges on the quality of the speech recognition.

In this paper, we propose the use of generative adversarial network for singing voice conversion, that doesn’t require a speech recognition system. Generative Adversarial Network (GAN) is a generative model that can learn a complex relationship between source and target features through an adversarial process. Recently, GANs have been successfully used in many fields such as image-to-image translation [32], speech recognition [33], and speech enhancement [34]. Moreover, GAN-based models such as GAN+WaveNet [35], VAW-GAN [36] and CycleGAN [37], [38] have shown to be effective in speech voice conversion. As GAN performs well with a smaller amount of training data than other deep neural networks, we hope to achieve high quality converted singing without the need of large training data.

The main contributions of this paper include, 1) we propose a novel singing voice conversion framework, that is based on Generative Adversarial Networks, 2) by using GAN, we eliminate the need of any external process, such as speech recognition, and reduce the reliance on large amount of training data, and 3) we achieve high quality singing voice that outperforms other DNN-based techniques. To our best knowledge, this paper reports the first successful attempt to use Generative Adversarial Networks in singing voice conversion.

This paper is organized as follows: In Section 2, we describe the Generative Adversarial Networks. In Section 3, we present our novel singing voice conversion framework. We report the experiments in Section 4 and conclude in Section 5.

II. GENERATIVE ADVERSARIAL NETWORKS (GANs)

A generative adversarial network [27] consists of a generator and a discriminator $D(x; \theta_D)$, where $\theta_D$ is the model parameters for the discriminator. In this structure, generator basically serves as a mapping function from distribution of source to distribution of target. The posterior probability of an input $x$ being a natural data, can be obtained by taking the sigmoid function from the discriminator’s output, $1/(1 + exp(-D(x)))$. The discriminator is trained to make the posterior probability 1 for natural data and 0 for generated data, while the generator is trained to deceive the discriminator. The system model of a traditional Generative Adversarial Network is given in Figure 2.

Generative adversarial networks have recently been shown to be an effective training method and have become popular in many fields such as image generation [39], image synthesis [39], speech enhancement [34], language identification [40], and text-to-speech synthesis [41]. Moreover, GANs have been recently used for speech voice conversion [35], [42] and achieve remarkable performance in terms of voice quality and speaker similarity. More recently, GAN-based speech voice conversion techniques include VAW-GAN [36], CycleGAN [37], CycleGAN-VC2 [43] and STARGAN-VC [44] that achieve remarkable performance with nonparallel training data. In this paper, we propose to use GANs for high-quality singing voice conversion.

III. GANS FOR SINGING VOICE CONVERSION

In this section, we explain the technical steps of our proposed GAN-based singing voice conversion framework. With parallel training data, we use source and target singing spectral features as the input for the generative adversarial network, we therefore call the proposed framework as SiGAN.

Singing voice conversion is a challenging task as the modeling and the conversion of singing spectrum is not straightforward. Moreover, singing is a form of art, and any distortion on the converted singing voice cannot be tolerated. To achieve high quality singing voice, there have been some statistical methods that are based on GMM [11], [12], by using parallel data. The furtherance in deep learning has a positive impact in many fields, that also include speech synthesis and voice conversion. However, deep neural networks have not
been well explored for singing voice conversion. To our best knowledge, this paper is the first to propose a GAN-based singing voice conversion framework.

SINGAN shares similar motivation with [30] regarding the use of deep learning in singing voice conversion. However, it differs from [30] in many ways, for example: 1) The fundamental difference is that we train a generative adversarial network to learn a mapping between source and target singers, while [30] uses DBLSTM to learn this mapping; 2) With SINGAN, we study a one-to-one singing voice conversion system with a small parallel training set, while [31] studies a many-to-one conversion with a large amount of non-parallel training data. 3) SINGAN does not hinge on automatic speech recognition (ASR) performance, while the approach in [30] performs mapping in between spectral features and PPGs, that are the intermediate results of an ASR system.

A. Training Phase

The training phase of the SINGAN framework is given in Figure 3. 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.

We propose to use GANs to learn the essential differences between the source singing and the original target singing through a discriminative process. Our GAN structure consists of two DNNs, that are iteratively updated by minibatch stochastic gradient descent. 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.

B. Run-time Conversion Phase

The run-time conversion phase of SINGAN is given in Figure 4. The run-time conversion phase also has 3 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, during the run-time conversion phase, we only convert the spectral features with the trained generative adversarial network. Previous studies [28] 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. In this paper, we do not perform F0 conversion for intra-gender SVC experiments. 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.

IV. EXPERIMENTS

In this section, we evaluate the effectiveness of the proposed singing voice conversion algorithm in terms of spectral feature conversion. We conduct both objective and subjective evaluations to compare the proposed SINGAN with a traditional DNN-based voice conversion framework.

A. Experimental Conditions

We conduct experiments on NUS Sung and Spoken Lyrics Corpus (NUS-48E corpus) [45] to assess the performance of the proposed SINGAN framework. The corpus consists of
audio recordings of the sung and spoken lyrics of 48 English songs by 12 professional singers. To assess the effect of limited data, we conduct experiments with 3, 5 and 7 source-target singing pairs. We use the WORLD vocoder [46] for feature analysis and synthesis. We extracted 34 Mel-cepstral coefficients (MCEPs), logarithmic fundamental frequency (log F0), and aperiodicities (APs) every 5 ms by using the WORLD analyzer. For preprocessing, we 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.

The proposed SINGAN is used to convert MCEPs (Q = 34 + 1 dimensions including 0th coefficient). Therefore, the objective of our experiments is to analyze the quality of the converted MCEPs. We directly copy the aperiodicity from source speaker. For intra-gender experiments, we do not perform F0 conversion, while for inter-gender 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. Dynamic time warping was used to align total frame lengths of the input and output speech parameters.

The proposed SINGAN structure consists of two DNNs, that are iteratively updated by minibatch stochastic gradient descent. In the experiments, we construct DNNs for male-to-male and female-to-male 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.

### B. Singing Voice Conversion with DNNs as a Baseline

As a baseline, we choose to use deep neural network (DNN) approach [47] to singing voice conversion. Our aim is to find a mapping between source and target singers by using parallel training data. Similar to that of SINGAN, we extracted 34 Mel-cepstral coefficients (MCEPs), logarithmic fundamental frequency (log F0), and aperiodicities (APs) every 5 ms by using the WORLD analyzer. We then normalized the source and target MCEPs to zero-mean and unit variance by using the statistics of the training sets. 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.

In the experiments, we constructed DNNs for male-to-male conversion and female-to-male conversion. The hidden layers of the DNN have 3 * 512 units. We note that DNN-based approach has been widely used as a baseline for GAN-based speech synthesis and voice conversion frameworks [41].

#### C. Objective Evaluation

We adopt the Mel-cepstral distortion (MCD) [28] between (1) the MCCs of source singer’s natural singing and the converted MCCs, that is denoted as $MCD_{source}$ and (2) the MCCs of target singer’s natural singing and the converted MCCs, that is denoted as $MCD_{cv}$. MCD values are calculated as follows:

$$MCD_{cv} = 10 \log_{10} \left( \frac{1}{35} \sum_{m=1}^{35} (c_{m} - c_{m,cv})^2 \right)$$

(1)

$$MCD_{source} = 10 \log_{10} \left( \frac{1}{35} \sum_{m=1}^{35} (c_{s,m} - c_{s,cv,m})^2 \right)$$

(2)

<table>
<thead>
<tr>
<th>Frameworks</th>
<th>Gender Information</th>
<th>Training Data (# song pairs)</th>
<th>MCD (cv)</th>
<th>MCD [source]</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN-based Baseline</td>
<td>male-to-male</td>
<td>3</td>
<td>5.15</td>
<td>6.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>5.98</td>
<td>6.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>5.73</td>
<td>6.49</td>
</tr>
<tr>
<td>SINGAN</td>
<td>male-to-male</td>
<td>3</td>
<td>5.81</td>
<td>6.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>5.52</td>
<td>6.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>5.36</td>
<td>6.03</td>
</tr>
<tr>
<td>DNN-based Baseline</td>
<td>female-to-male</td>
<td>3</td>
<td>6.32</td>
<td>7.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>6.12</td>
<td>6.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>5.89</td>
<td>6.65</td>
</tr>
<tr>
<td>SINGAN</td>
<td>female-to-male</td>
<td>3</td>
<td>5.97</td>
<td>6.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>5.74</td>
<td>6.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>5.53</td>
<td>6.12</td>
</tr>
</tbody>
</table>

Table 1: A summary of the comparison between the DNN-based baseline framework, and the proposed SINGAN framework. We conduct experiments with 3, 5 and 7 source-target song pairs to show the effect of limited training data. In all experiments, parallel data have been used during training.
where $c_{cv}(m)$, $c_{t}(m)$ and $c_{s}(m)$ are the $m^{th}$ coefficients of the converted singing MCCs, target singing MCCs, and source singing MCCs, respectively. We calculate the MCD values frame-by-frame over all the paired frames in the test set, and report the average MCD values. We note that a lower MCD value indicates smaller spectral distortion.

Moreover, we report the signal to noise ratio (SNR) [35] between converted singing waveform and target singing waveform. SNR values are calculated as follows:

$$SNR[dB] = 10 \log_{10} \left( \frac{\sum_{n=1}^{N} x(n)^2}{\sum_{n=1}^{N} (x(n) - y(n))^2} \right)$$  \hspace{1cm} (3)$$

where $x(n)$ is the converted singing waveform, and $y(n)$ is the target singing waveform at time $n$. The objective evaluation results are shown in Table 1 and Figure 5.

In Table 1, we report the $MCD_{[source]}$ and $MCD_{[cv]}$ under different training settings. We would like to compare the proposed SINGAN and the DNN-based baseline framework in terms of training data size and gender of the singers. Firstly, we see that SINGAN outperforms the DNN-based approach by achieving lower $MCD_{[cv]}$ and $MCD_{[source]}$ in all cases. Secondly, the proposed SINGAN achieves better performance in intra-gender singing voice conversion, that is also consistent with the DNN-based approach. Thirdly, we observe that $MCD_{[cv]}$ is always lower than $MCD_{[source]}$. The results suggest that SINGAN framework generates a singing spectrum, that is more similar to the original target singer. Last but not least, we would like to note that the proposed SINGAN can work remarkably well with limited amount of parallel data and outperforms the baseline in all settings. For example, SINGAN (male-to-male) with 3 song pairs achieves the $MCD_{[cv]}$ value of 5.81, while baseline DNN (male-to-male) achieves the $MCD_{[cv]}$ value of 5.98 with 5 song pairs.

In Figure 5, we report the SNR values under different training settings. We can see that SINGAN always outperforms the DNN-based baseline framework by achieving higher SNR, that is also consistent with the previous experiments. We can also see that intra-gender singing voice conversion achieves better performance than inter-gender singing voice conversion, both in DNN-based approach and SINGAN.

D. Subjective Evaluation

We conduct two listening experiments to assess the performance of the proposed SINGAN for singing voice conversion, in terms of voice quality and speaker similarity. 20 subjects participated in all the listening tests. Each subject listens to 30 converted singing samples.

Firstly, we evaluate the sound quality of the converted voices with mean opinion score (MOS), that is reported in Figure 6. 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 compare SINGAN and DNN in terms of singing voice conversion performance. We use 7 source target song pairs to train the SINGAN and the baseline DNN framework. As can be seen, SINGAN outperforms the DNN-based baseline framework by achieving the MOS value of 3.89 ±0.11.

We further conduct preference test, that is reported in Figure 7, to compare SINGAN with DNN baseline, in terms of speaker similarity. To show the capability of our proposed framework under limited data, we use 7 song pairs for DNN training, while we only use 5 song pairs for SINGAN training. We show that SINGAN outperforms the DNN-based approach in terms of speaker similarity, even with less training data, as it is chosen as the better sample for 66.8 ± 2.2 percent of the time. Singing samples can be found in the following link: https://sites.google.com/view/berraksisman/.

V. CONCLUSION

In this paper, we propose a novel singing voice conversion framework, that is based on generative adversarial networks. The proposed approach performs remarkably well with very limited parallel training data from both singers. In the experiments, we outperform the baseline and achieve high-quality singing voice. We believe that proposed approach produce good results and can even serve as baseline SVC framework in the future.

We have also tried applying generative adversarial networks for singing voice conversion with nonparallel data and have obtained some good preliminary results. More investigation on the parallel-data-free singing voice conversion will be conducted in the future.
ACKNOWLEDGMENT
This research is supported by Programmatic grant no. A18A2b0046 from the Singapore Governments Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain). This research is also supported by the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG-100E-2018-006). Berrak Sisman is funded by A*STAR Graduate Academy under Singapore International Graduate Award (SINGA).

REFERENCES


