Change Your Singer: A Transfer Learning Generative Adversarial Framework for Song to Song Conversion

Rema Daher*
Dept. of Mechanical Engineering
American University of Beirut
Beirut, Lebanon
rgd05@mail.aub.edu

Mohammad Kassem Zein*
Dept. of Mechanical Engineering
American University of Beirut
Beirut, Lebanon
mhk50@mail.aub.edu

Julia El Zini
Dept. of Electrical and Computer Engineering
American University of Beirut
Beirut, Lebanon
jwe04@mail.aub.edu

Mariette Awad
Dept. of Electrical and Computer Engineering
American University of Beirut
Beirut, Lebanon
ma162@aub.edu.lb

Daniel Asmar
Dept. of Mechanical Engineering
American University of Beirut
Beirut, Lebanon
da20@aub.edu.lb

***Abstract*—*Have you ever wondered how a song might sound if performed by a different artist?* In this work, we propose SCM-GAN, an end-to-end non-parallel song conversion system powered by generative adversarial and transfer learning, which allows users to listen to a selected target singer singing *any* song. SCM-GAN first separates songs into vocals and instrumental music using a U-Net network, then converts the vocal segments to the target singer using advanced CycleGAN-VC, before merging the converted vocals with their corresponding background music. SCM-GAN is first initialized with feature representations learned from a state-of-the-art voice-to-voice conversion and then trained on a dataset of non-parallel songs. After that, SCM-GAN is evaluated against a set of metrics including global variance GV and modulation spectra MS on the 24 Mel-cepstral coefficients (MCEPs). Transfer learning improves the GV by 35% and the MS by 13% on average. A subjective comparison is conducted to test the output's similarity to the target singer and its naturalness. Results show that the SCM-GAN's similarity between its output and the target reaches 69%, and its naturalness reaches 54%.**

***Index Terms*—Generative Adversarial Networks, Song to Song Conversion, Voice to Voice, Transfer Learning**

I. Introduction

Voice-to-voice conversion is the process of converting a speech spoken by a particular speaker to another selected target speaker. Prior work on voice-to-voice using deep learning utilized sequence-to-sequence voice conversion [1], and phoneme-based linear mapping functions [2] to provide an end-to-end voice-to-voice solution. Recently, generative adversarial networks (GANs) have shown their success in natural language processing [3], and image and video synthesis [4]. Given the requirement of generating a voice that mimics a particular data distribution, systems based on GANs [5] showed promising results in voice-to-voice conversions.

Song-to-song systems are a particular case of the voice-to-voice problem, and attempt to change existing songs by incorporating the voice of a user-selected artist. Such systems have many practical applications. For instance, music applications can integrate novel features that allow users to listen to any song by the voice of their favorite singer. Additionally, users can pretend on social media platforms to sing a song by replacing the voice of the original singer by their own voice.

Given that speech and music encode distinct sorts of information differently, their acoustical features are fundamentally dissimilar [6]. For instance, fundamental frequencies, temporal regularities and quantization, short silences, steady and varying formants, and transient spectral details significantly vary between speech and music. This makes speech recognition challenging when there is even a modest level of background music [6]. Existing song-to-song systems only focus on achieving the singing voice conversion without developing stand-alone end-to-end systems that perform well when background music is inputted along the vocals [7].

In this work, we propose a novel end-to-end system powered by generative adversarial networks and transfer learning, and which replaces the original singer of a song by any desired performer. The long term objective of the proposed system aims at enabling developers to build a commercial application for the aforementioned purpose. Our model, **S**plit-**C**onvert-**M**erge using Cycle**GAN**, SCM-GAN, first takes advantage of the U-Net [8] to split vocals from the background music, then trains an instance of Voice Converter CycleGANs [9] on a set of in-house collected songs, not necessarily parallel. Finally, SCM-GAN merges the converted singing voice with the background music to achieve the song-to-song conversion. Moreover, we utilize acoustic features learnt in the voice converter CycleGAN of [9] to efficiently train SCM-GAN using

*Authors with equal contribution

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

transfer learning. We show the importance of the suggested system through objective and subjective evaluation. The results show that SCM-GAN successfully converts a song to a target singer song with high resemblance to ground truth. Transfer learning improves the average GV and MS by 35% and 13% respectively and the splitting scheme increases the subjective evaluation scores by 73% and 26% on the naturalness and similarity of the converted song. The contributions of this work include: (1) an end to end song-to-song conversion approach which (2) combines two mostly-unrelated machine learning tasks with co-dependent results. The purpose of this combination is to (3) split and reconstruct songs as well as (4) transfer knowledge from voice-to-voice models.

The remainder of the paper is organized as follows: first, related work is reported in Section II, then Section III describes the system Methodology before Section IV reports the results of the conducted experiments. Section V concludes with final remarks.

II. Related Work

Converting a specific speaker's voice to a target voice is a topic of interest for many researchers. For instance, [10] presented a voice conversion technique by introducing a STRAIGHT mixed excitation [11] to Maximum Likelihood Estimation (MLE) with a Gaussian Mixture model (GMM). However, in their work they focused on the quality of the converted voice rather than the conversion accuracy. Another approach of voice conversion was introduced by [12]. Their system involved mapping the spectral features of a source speaker to that of a target speaker using Artificial Neural Network (ANN); they proved that the mapping capabilities of (ANN) perform better transformation and produce better quality voices than GMMs.

Deep Neural Networks (DNNs) were also used in voice conversion systems [13]. However, the problem with conventional DNN frame-based methods is that they do not capture the temporal dependencies of a speech sequence. To tackle this problem, [14] proposed a Deep Bidirectional Long Short-Term Memory based Recurrent Neural Network (DBLSTM-RNN) architecture to model the long-range context-dependencies in the acoustic trajectory between the source and the target voice. Recently, Kaneko and Kameoka in [9] proposed a novel voice converter (CycleGAN-VC) that relies on cycle-consistent adversarial networks (CycleGAN).

For the conversion of voices that are sung, [15] relied on statistical modeling of the speakers' timbre space using a GMM in order to define a time-continuous mapping from the feaures of the the source speaker to the target. Moreover, [16] used direct waveform modification based on the spectrum differential to achieve the conversion of the singing voice. Then, they extended it to restore the global variance of the converted spectral parameter trajectory to avoid oversmoothing at unvoiced frames in [7].

So far, the attempts of conversion of singing voice have relied on methods used for voice-to-voice conversion. However, to the best of our knowledge, the state of the art method in voice conversion, CycleGAN-VC [9], has not yet been implemented for the conversion of singing voice. In this paper, we propose an end-to-end system SCM-GAN that employs CycleGAN-VC [9] along with a deep U-Net [8] to achieve song-to-song conversion.

III. Methodology

In this work, we propose SCM-GAN, an end-to-end system that converts songs from the voice of *any* singer to that of a *specific fixed* target singer $\mathcal{S}$ without altering the background music. For this purpose, CycleGAN-VC$^{\text{voc}}$, a deep CycleGAN converter that is trained on instances of the form $(voc_{i,A}, voc_{i,\mathcal{S}})$ where $voc_{i,\mathcal{S}}$ is the i^{th} vocals segment sung by the singer S. To be able to maintain the background music, only vocals are fed into CycleGAN-VC$^{\text{voc}}$ after being separated from the background music by a deep U-Net [8].

The overall workflow of SCM-GAN is shown in Fig. 1. First, the song is fed into a pre-trained U-Net model [8], which separates songs into vocals and background music. Then, the vocals are inputted into the CycleGAN-VC$^{\text{voc}}$, which converts them into the voice of $\mathcal{S}$ using transfer learning. Finally, a merging scheme is used to overlay the converted output with the saved background music from the separation phase. The system is composed of four main components including: (1) the vocals-music separation with a pre-trained U-Net [8] (2) the vocals conversion provided by CycleGAN-VC, (3) the knowledge transferred from a CycleGAN-VC network trained on *speech* and (4) the merging scheme. Four different notations will be used throughout the paper to distinguish between CycleGAN-VC implementations: (1) CycleGAN-VC$^{\text{sp}}$ trained on speech, 2) CycleGAN-VC$^{\text{voc}}$ trained on vocals using transfer learning, 3) CycleGAN-VC$^{\text{voc}}_{\text{scratch}}$ trained on vocals from scratch and 4) CycleGAN-VC$^{\text{voc+music}}$ trained on song (vocals and background music) using transfer learning.

A. Music-Vocals Separation

A U-Net [8] is implemented as an encoder-decoder fully connected convolutional neural network to separate the background music from the vocals by operating exclusively on the magnitude of audio spectrograms. Specifically, the U-Net implements two decoders: one for the instrumental music and another one for the vocals. The audio signal for both components (instrumental/vocal) is recompiled as follows: the magnitude component of the signal is reconstructed by applying the output mask of each decoder to the magnitude of the original spectrum; and its phase component is that of the original spectrum without any modifications.

B. Vocals Conversion With CycleGAN-VC^{voc}

In this work, vocals are converted by CycleGAN-VC$^{\text{voc}}$ (trained on vocals) that has the same underlying architecture as in [9]. The CycleGAN-VC architecture modifies that of cycleGAN [17], adding to it identity mapping loss [18] and a gated CNN [19] that can represent sequential and

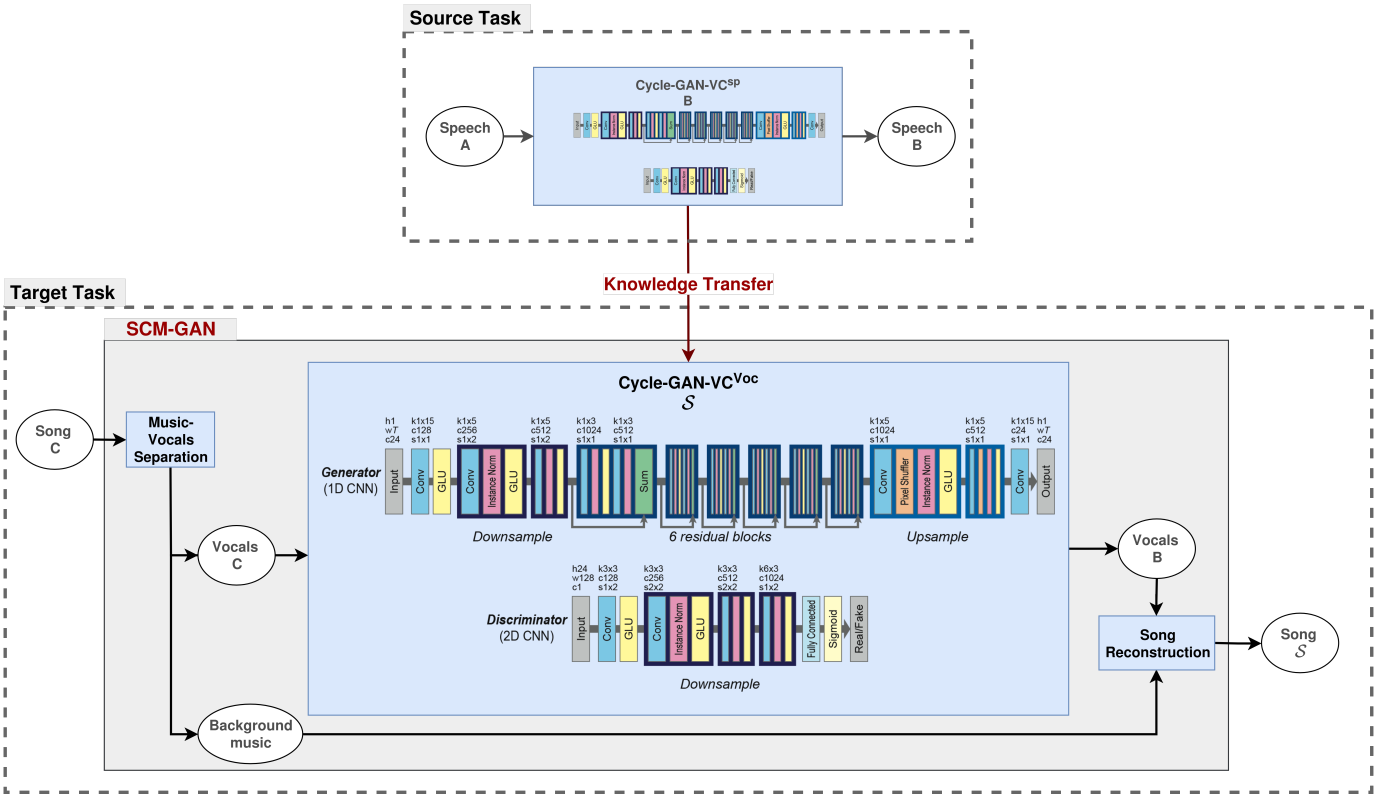

Fig. 1. SCM-GAN system overview. [9] provides details of the adopted cycle-GAN architecture

hierarchical features of speech, and generate state-of-the-art speech output [1]. The superior results are achieved because of the networks' structure, which include gated linear units (GLUs) that act as data driven activation functions:

$$H_{l+1} = (H_l * W_l + b_l) \otimes \sigma(H_l * V_l + c_l),$$

where H_{l+1} and H_l are the $l+1$ and l layer outputs respectively. In addition, W_l, b_l, V_l, and c_l represent the parameters of the model and σ represents the *sigmoid* function. Here, $\otimes$ is the element-wise product.

In the CycleGAN-VC architecture, three losses are utilized, including an adversarial loss, a cycle-consistency loss, and an identity-mapping loss [18]. First, we denote the mapping from the source $x \in X$ to the target $y \in Y$ as $G_{X\rightarrow Y}$ and the reciprocal as $G_{Y\rightarrow X}$. Then, the adversarial loss can be described as the difference between the distribution of converted data $P_{G_{X\rightarrow Y}}(x)$ and their corresponding actual training output distribution $P_{Data}(y)$. In order to reduce this difference and deceive the discriminator D_Y by getting an output close to the target output, the following objective function is minimized:

$$L_{adv}(G_{X\rightarrow Y}, D_Y) = E_{y\sim P_{Data}(y)}[logD_Y(y)]+ \\ E_{x\sim P_{Data}(x)}[log(1 - D_Y(G_{X\rightarrow Y}(x)))]$$

Conversely, D_Y maximizes this loss to avoid being deceived. Cycle-consistency loss attempts to keep the contextual information between the input and the converted output consistent using the following objective function:

$$L_{cyc}(G_{X\rightarrow Y}, G_{Y\rightarrow X}) = \\ E_{x\sim P_{Data}(x)}[\parallel G_{Y\rightarrow X}(G_{X\rightarrow Y}(x)) - x \parallel_1]+ \\ E_{y\sim P_{Data}(y)}[\parallel G_{X\rightarrow Y}(G_{Y\rightarrow X}(y)) - y \parallel_1]$$

Identity-mapping loss is used to preserve composition and linguistic information between input and output. This loss is defined as follows:

$$L_{id}(G_{X\rightarrow Y}, G_{Y\rightarrow X}) = E_{y\sim P_{Data}(y)}[\parallel G_{X\rightarrow Y}(y) - y \parallel_1]+ \\ E_{x\sim P_{Data}(x)}[\parallel G_{Y\rightarrow X}(x) - x \parallel_1]$$

Using an inverse adversarial loss $L_{adv}(G_{Y\rightarrow X}, D_X)$, the full objective function can be expressed as:

$$L_{full} = L_{adv}(G_{X\rightarrow Y}, D_Y) + L_{adv}(G_{Y\rightarrow X}, D_X)+ \\ \lambda_{cyc}L_{cyc}(G_{X\rightarrow Y}, G_{Y\rightarrow X}) + \lambda_{id}L_{id}(G_{X\rightarrow Y}, G_{Y\rightarrow X})$$

where λ_{cyc} and λ_{id} are trade-off parameters for their corresponding losses. This objective function allows the model to learn the mapping from a source singer to a target one chosen by the user.

C. Knowledge Transfer from CycleGAN-VC^{sp}

CycleGAN-VC$^{\text{sp}}$ [9] is trained on instances of the form $(\text{sp}_{i,A}, \text{sp}_{i,B})$ where $\text{sp}_{i,S}$ is the i^{th} speech training instance spoken by S. In order to speed up the training of CycleGAN-VC$^{\text{voc}}$, voice feature representation learnt in CycleGAN-VC$^{\text{sp}}$ is used to initialize the training. It is worth mentioning that the target speaker in CycleGAN-VC$^{\text{sp}}$ is different than $\mathcal{S}$, the target speaker of our CycleGAN-VC$^{\text{voc}}$.

D. Song Reconstruction

Since the proposed pipeline maintains the temporal characteristic of the input audio after the splitting and converting step, it is enough to just overlay the background music with the output from the model. In order to overlay the converted vocals with their corresponding instrumental music, they are both segmented via an analysis of signal onsets and offsets as in [20]. Segments are then integrated at a coarse scale and at a finer scale by locating accurate onset and offset positions for segments as in [21].

IV. Experiments

To assess the efficiency of SCM-GAN, we performed several experiments in which a performer is replaced by another singing the identical song. Since singing voice conversion methods have relied on voice-to-voice conversion, we will be comparing SCM-GAN to the state of the art method in voice-to-voice conversion (CycleGAN-VC). Given that no public dataset exists that includes two voices singing the same part with background music, we developed our own.

A. Dataset

We collected a dataset of non-parallel aligned song segments of $3sec$ each on average to be consistent with the Voice Conversion Challenge 2016 (VCC 2016) dataset [22] used by [9]. For every instance of a singer A, a corresponding instance is created with the target singer $\mathcal{S}$ singing the same lyrics at a different time frame. Particularly, 228 training instances were created with Samantha Harvey as singer A and Ed Sheeran as target singer $\mathcal{S}$. As for the testing data, 15 instances of $10secs$ each (for better subjective and objective evaluation) were collected. These instances include songs by singer A singing her own songs as opposed to singing singer B's songs. In addition, songs from 5 singers different from A were also included in the testing data including: Beyonce, Bea Miller, Diamond White, Nicole Cross, and Chelsea FreeCoustic. This dataset will be made publicly available for further research and improvements in the field.

After splitting the data into vocals and background music, the vocals were then pre-processed by downsampling the data to 16 kHz. Afterwards, at every 5 ms the data is transformed into MCEPs, aperiodicities (APs), and logarithmic fundamental frequency $log(F_0)$, using a speech synthesis system WORLD [23]. Then, a normalized logarithm Gaussian transformation, [2], is applied on F_0 as in [9].

B. Training

After the separated vocals are preprocessed, voice-to-voice inter-gender weights ($SF1 - TF2$) from [9] are loaded into CycleGAN-VC, which is then fine-tuned on 1000 epochs using our data. The choice of the number of epochs was chosen to be 1 since the losses converged after that. Fine-tuning can be used in this case since the task the model has already learnt (voice-to-voice) is similar to the new task it is about to learn (song-to-song). The reason behind fine-tuning is that much less epochs and processing time are needed than that needed to train a model from scratch.

C. Objective Evaluation

To properly assess our proposed system, we evaluated the quality of the converted feature vector (MCEPs). Specifically, we focused in our experiments on analyzing two associated metrics: Modulation Spectrum (MS) [24] and Global Variance (GV) [25]. To test the importance of transfer learning, we compared our system CycleGAN-VC$^{\text{voc}}$ to CycleGAN-VC$^{\text{voc}}_{\text{scratch}}$ (trained on vocals from scratch), a model with same architecture but has not been pretrained with the knowledge from CycleGAN-VC$^{\text{sp}}$. Figs. 2 and 3 present the comparison of GV and MS respectively between the two models compared to the ground truth. The root mean squared errors (RMSEs) between the models and the target are calculated and summarized in Table I showing that the RMSEs of CycleGAN-VC$^{\text{voc}}$ with the target on the basis of GV and MS are smaller than those of CycleGAN-VC$^{\text{voc}}_{\text{scratch}}$. Consequently, transfer learning improves the average GV and MS by 35% and 13% respectively.

To further validate the previous results, the change in CycleGAN-VC$^{\text{voc}}$'s losses throughout the training is compared to that of CycleGAN-VC$^{\text{voc}}_{\text{scratch}}$. It is worth mentioning that losses don't converge to zero, since GANs are used, which are nothing but a play on losses between the generator and the discriminator. Fig. 4 shows that the former outperformed the latter in terms of jump start performance and the final loss.

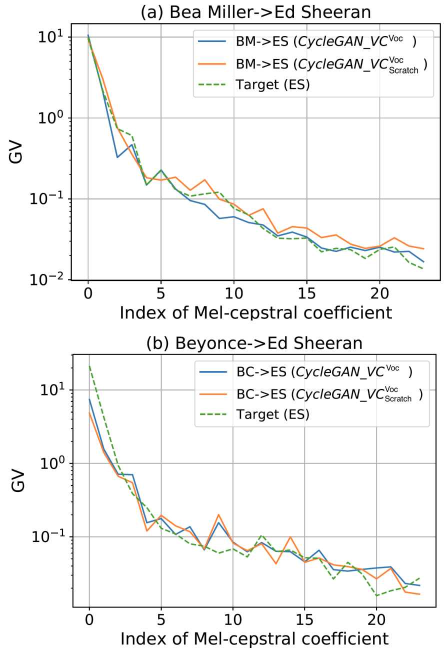

Fig. 2. GV of CycleGAN-VC$^{\text{voc}}$ and CycleGAN-VC$^{\text{voc}}_{\text{scratch}}$ on two song segments (best seen in color)

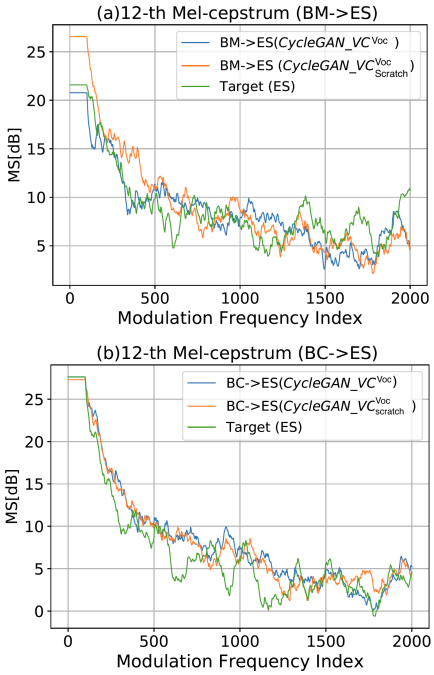

Fig. 3. MS of CycleGAN-VC$^{\text{voc}}$ and CycleGAN-VC$^{\text{voc}}_{\text{scratch}}$ on two song segments (best seen in color)

TABLE I
COMPARISON OF AVERAGE RMSE BETWEEN TARGET AND CONVERTED INSTANCES IN TERMS OF GV AND MS. SMALLER VALUES INDICATE RESEMBLANCE TO TARGET.

Method	GV	MS
CycleGAN-VC$^{\text{voc}}$	**1.735 ± 1.076**	**6.833 ± 0.373**
CycleGAN-VC$^{\text{voc}}_{\text{scratch}}$	2.696 ± 2.386	7.922 ± 0.27

D. Subjective Evaluation

We found it difficult to objectively test the effect of splitting the background music from vocals because the target ground truth song does not have the same background music nor the same pace. Hence, we subjectively evaluated the importance of splitting background music before being inputted to CycleGAN-VC$^{\text{voc}}$, and we trained CycleGAN-VC$^{\text{voc+music}}$ (trained on song, vocals and background music, using transfer learning), on full song segments without splitting, then compared it to CycleGAN-VC$^{\text{voc}}$.

To perform the subjective tests, we prepared a survey for song evaluation using five $10sec$ song segments that were converted using two models: CycleGAN-VC$^{\text{voc}}$ and CycleGAN-VC$^{\text{voc+music}}$. The survey was filled by twenty test subjects of random gender, age, and musical background. The survey was conducted according to a ranking system on similarity basis ranging from 1 (similar to original singer) to 5 (similar to target speaker). Furthermore, naturalness was also included in the survey with a $1-5$ score ranging from not natural to very natural. The data from the survey was then analyzed using mean opinion score (MOS) test. The results in Table II show that the output of our system CycleGAN-VC$^{\text{voc}}$ is closer to the target (reaching a 69% similarity to the target) than CycleGAN-VC$^{\text{voc+music}}$ (reaching only a 55% similarity to the target) by 26%. This was accompanied with an increase of 73% in the degree of naturalness on average with our system CycleGAN-VC$^{\text{voc}}$, and CycleGAN-VC$^{\text{voc+music}}$ having 54% and 31% naturalness respectively. We demonstrated that our system CycleGAN-VC$^{\text{voc}}$ has higher MOS than CycleGAN-VC$^{\text{voc+music}}$. Particularly, we confirmed that data with background music has an adverse effect on the performance of the conversion model as expected, and adding a separation model to the pipeline had valuable implications on the output.

TABLE II
COMPARISON OF MOS OF NATURALNESS AND SIMILARITY BETWEEN TARGET AND CONVERTED. LARGER VALUES INDICATE RESEMBLANCE TO TARGET AND HIGHER NATURALNESS.

Method	Naturalness	Similarity
CycleGAN-VC$^{\text{voc}}$	**2.68**	**3.46**
CycleGAN-VC$^{\text{voc+music}}$	1.55	2.75

E. Limitations

The analyzed results are also coupled with limitations that we will address in future work. The encountered limitations include the modest size of the dataset that had to be developed manually. That is, there is no ready dataset that includes two voices singing the same part with background music. Other limitations come from the drawbacks of using a subjective survey-based evaluation method, which may include dishonest answers, missing data, social desirability bias, unconscientious responses, and others.

V. CONCLUSION

In this paper, we presented our novel end to end framework, SCM-GAN that successfully transformed songs to be performed by a target singer using an in-house collected dataset of non parallel songs. This was achieved by utilizing U-Net, Generative Adversarial Networks, and encoder-decoder architectures to first separate the songs from their background music, convert them, and then reconstruct the target song. The results were evaluated on the basis of the global variance and modulation spectrum of their corresponding Mel-spectrum coefficients which showed that transfer learning improves the performance of SCM-GAN by 35% in the global variance. The naturalness and similarity to the ground truth of the system output was evaluated with a subjective survey that shows the SCM-GAN's output having 69% similarity to the ground truth and 54% naturalness. The encouraging results of our model SCM-GAN pave the way for an expansion into models that easily adapt to different target singers and languages through advanced forms of transfer learning.

ACKNOWLEDGMENT

We would like to thank the University Research Board of the American University of Beirut for funding this work, as well as Revotonix L.L.C for providing Amazon Web Services (AWS) credits for training the model.

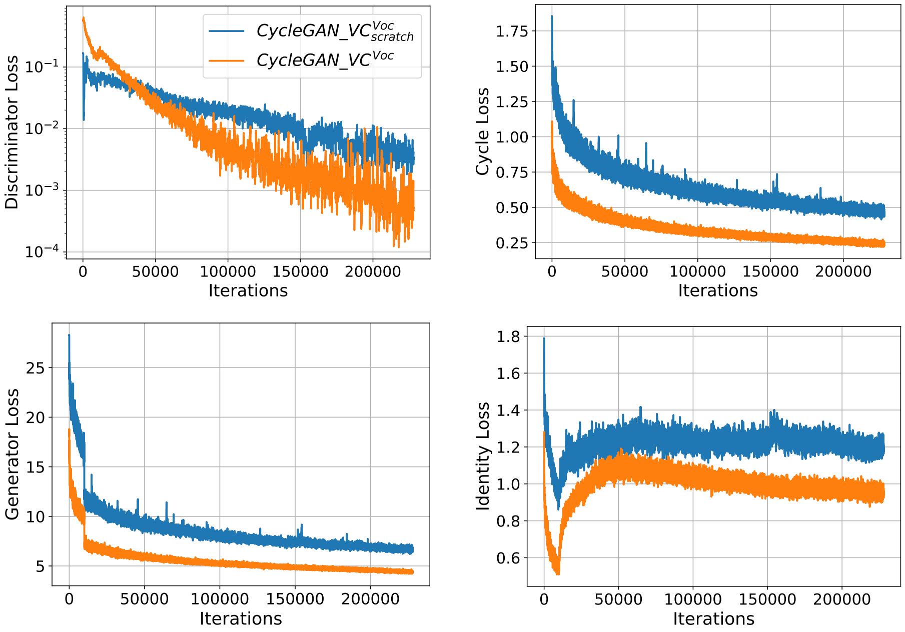

Fig. 4. Comparison of losses for CycleGAN-VC$^{\text{voc}}$ and CycleGAN-VC$^{\text{voc}}_{\text{scratch}}$(best seen in color)

References

[1] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks." in *INTERSPEECH*, 2017, pp. 1283–1287.

[2] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through phoneme-based linear mapping functions with straight for mandarin," in *Fourth FSKD International Conference*, vol. 4. IEEE, 2007, pp. 410–414.

[3] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward controlled generation of text," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1587–1596.

[4] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE Conference CVPR*, 2018, pp. 1316–1324.

[5] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High fidelity speech synthesis with adversarial networks," *arXiv preprint arXiv:1909.11646*, 2019.

[6] J. Wolfe, "Speech and music, acoustics and coding, and what music might be 'for'," in *Proc. 7th International Conference on Music Perception and Cognition*, 2002, pp. 10–13.

[7] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion based on direct waveform modification with global variance," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *18th International Society for Music Information Retrieval Conference*, 2017, pp. 23–27.

[9] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *26th EUSIPCO*, 09 2018, pp. 2100–2104.

[10] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on gmm with straight mixed excitation," in *Conference of the International Speech Communication Association*, September 2006, pp. 2266–2269.

[11] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

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

[13] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using deep belief nets." in *Interspeech*, 2013, pp. 369–372.

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

[15] F. Villavicencio and J. Bonada, "Applying voice conversion to concatenative singing-voice synthesis," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[16] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[18] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *5th International Conference on Learning Representations, ICLR*, Toulon, France, April 2017.

[19] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 933–941.

[20] G. Hu and D. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 396–405, 2007.

[21] J.-p. S. James Robert, Marc Webbie *et al.*, "Pydub," Online, 2011. [Online]. Available: http://pydub.com

[22] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016." in *Interspeech*, 2016, pp. 1632–1636.

[23] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[24] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A

postfilter to modify the modulation spectrum in hmm-based speech synthesis," in *2014 IEEE International ICASSP*. IEEE, 2014, pp. 290–294.

[25] 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.