Effectiveness of Transfer Learning on Singing Voice Conversion in the Presence of Background Music

Divyesh G. Rajpura, Jui Shah, Maitreya Patel, Harshit Malaviya, Kirtana Phatnani, Hemant A. Patil
Speech Research Lab,
Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT),
Gandhinagar, Gujarat, India
{divyesh_rajpura, jui_shah, maitreya_patel, harshit_malaviya, kirtana_phatnani, hemant_patil} @daiict.ac.in

***Abstract*—Singing voice conversion (SVC) is a task of converting the perception of the source speaker's identity to the target speaker without changing lyrics and rhythm. Recent approaches in traditional voice conversion involve the use of the generative models, such as Variational Autoencoders (VAE), and Generative Adversarial Networks (GANs). However, in the case of SVC, GANs are not explored much. The only system that has been proposed in the literature uses traditional GAN on the parallel data. The parallel data collection for real scenarios (with the same background music) is not feasible. Moreover, in the presence of background music, SVC is one of the most challenging tasks as it involves the source separation of vocals from the inputs, which will have some noise. Therefore, in this paper, we propose transfer learning, and fine-tuning-based Cycle consistent GAN (CycleGAN) model for non-parallel SVC, where music source separation is done using Deep Attractor Network (DANet). We designed seven different possible systems to identify the best possible combination of transfer learning and fine-tuning. Here, we use a more challenging database, MUSDB18, as our primary dataset, and we also use the NUS-48E database to pre-train CycleGAN. We perform extensive analysis via objective and subjective measures and report that with a 4.14 MOS score out of 5 for naturalness, the CycleGAN model pre-trained on NUS-48E corpus performs the best compared to the other systems described in the paper.**

Keywords: Singing Voice Conversion, Deep Attractor Network, Transfer Learning, CycleGAN.

I. Introduction

Music is a form of art, which is derived from an organized structure of sounds, varying in frequency, however, played together in a symphony. In addition, speech is the most powerful and natural form of communication between humans. Moreover, via singing voice, one can express more skillfully. A singer can express more varieties of expressions using rhythm, notes, temporal dynamics, and, more importantly, linguistic content. Moreover, by changing the voice characteristics (such as voice timbre, vibrato, intonation, etc.), singers can also express themselves in many different emotions [1]. However, due to the physiological constraints during speech production, it is very difficult to change someone's voice with large variations [2]. The singing voice was studied deeply with acoustic and signal processing perspectives, and it led to one of the first few singing synthesizers, which were made using spectral and physical methods [3], [4].

Singing Voice Conversion (SVC) is the task of transforming the source singer's voice to sound like the voice of the target singer without changing the linguistic content, and the rhythm [1]. Current solutions do not pay much attention to SVC with background music despite the fact that SVC in the presence of background music has many applications, such as dubbing of the songs, singing voice synthesis, etc. Therefore, in this paper, we focus on SVC in the presence of background music. Here, background music can be composed of drums, piano, or any other instruments.

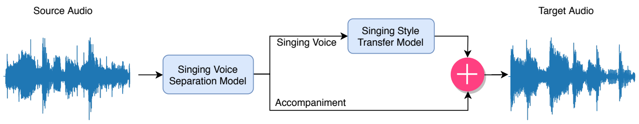

Fig. 1: Proposed singing style transfer framework.

In this paper, we approach the problem of SVC with background music via a two-step process. As shown in Fig. 1, the first step is to extract the vocals (i.e., linguistic content) from the songs using music separation, and the second involves voice conversion. Finally, we merge the converted singing voice with accompaniment extracted from the music separation. We focus on non-parallel data due to the unavailability of the parallel data.

With the advent of deep learning, state-of-the-art methods are continuously refined and tested for different speech technology problems. Blind Source Separation (BSS) used to separate various sources present in the audio mixture. In BSS, to separate different sources, the aim is to estimate a Time-Frequency (T-F) mask of each source given the T-F representation of the audio mixture. There have been many different deep learning-based approaches proposed for BSS recently, such as Deep Clustering (DPCL) [5], [6], Permutation Invariant Training (PIT) [7], and Deep Attractor Network (DANet) [8], [9], etc. DPCL was proposed for the more challenging task of speaker-independent speech separation. DANet uses a similar underlying concept as DPCL and modifies it to make an end-to-end BSS system. In [10], DANet has also been used for music separation, wherein they have tested model for separating same-class sources (i.e., separating different accompaniments), and between-class sources (i.e.,

978-1-7281-8895-9/20/$31.00 © 2020 IEEE

separating vocal *vs*. accompaniment). More details of DANet are specified in Section II-A. In the field of voice conversion, methods, such as Variational Autoencoders (VAEs), and Deep Bidirectional Long Short Term Memory (DBLSTM), have been proposed for non-parallel VC [11], [12]. A recently proposed method that outperforms the aforementioned is Generative Adversarial Network (GAN) [13]. GANs have shown remarkable success in parallel and non-parallel tasks, such as CycleGAN-VC, CycleGAN-VC2, and AdaGAN are state-of-the-art methods proposed for one-to-one non-parallel mapping [14]–[16]. Additionally, StarGAN-VC, and StarGAN-VC2 are proposed for many-to-many VC tasks [17], [18]. However, developing a non-parallel SVC is quite a challenging task compared to the parallel VC. Attempts have been made to develop many non-parallel SVC tasks. For example, the system with CNN-based encoder with Wavenet-based decoder, and RNN with DBLSTM structure have been recently proposed for the same [19], [20]. That being said, GAN-based methods have not been tested and validated for non-parallel SVC tasks. To the best of authors' knowledge, the only proposed system for SVC uses a traditional GAN-based approach [2], however, it only considers parallel data.

In this paper, we propose a system for SVC with background music, which extracts vocals from the songs via a DANet, and a style transfer system for non-parallel data with CycleGAN. Considering the limited non-parallel data available, we focus on the transfer learning and fine-tuning-based approach in this paper. To the best of authors' knowledge, this is the first study where transfer learning is being extensively tested for the SVC task. We pre-train DANet on the MUSDB18 [21] database, and CycleGAN via many different scenarios, which also includes pre-training, and fine-tuning on different datasets (either MUSDB18 [21], or NUS-48E [22]). After importing both the pre-trained model, we fine-tune CycleGAN according to our primary database, MUSDB18. This way, we generated a total of seven different possible scenarios, which involve transfer learning, and fine-tuning of the CycleGAN. We use various subjective and objective measures to validate each system. The key contributions of this paper are as follows:

- We extensively study the non-parallel SVC task using CycleGAN in the presence of different background music.
- We propose a novel approach of transfer learning and fine-tuning and analyze its effect on the SVC task.
- We perform SVC on the MUSDB18 database, which contains the reverberations of the singer's vocals, different pitches, and rhythms in a single song making it a difficult (in terms of complexity) database to work.

II. Proposed Framework

A. Deep Attractor Network (DANet) for Source Separation

Let us denote X as a mixture speech signal contains P number of sources, $x_1, x_2, ..., x_P$. Let $s_i \in \mathbb{R}^{F\times T}$, and $m_i \in \mathbb{R}^{F\times T}$ be a T-F representation, and ideal T-F mask of source i, where $i = 1, 2, ..., P$, respectively. Let $S \in \mathbb{R}^{F\times T}$ be a T-F representation of mixture speech signal, which is the sum of all P sources present in the mixture audio. Here, T represents the duration of utterance, and F represents the number of frequency channels. DANet tries to learn embedding for each T-F bin in D-dimensional space such that T-F bins from the same source are nearer to each other, and vice-versa [8]. T-F representation of mixture audio signal S serves as input to the network, which generates an embedding matrix, $E \in \mathbb{R}^{D\times FT}$, which contains D-dimensional embeddings for each T-F bin. Then, an attractor $a_i \in \mathbb{R}^{1\times D}$, which represents the centroid of source i in the embedding space is formed by calculating the weighted average of embeddings given the oracle mask, $m_i \in \mathbb{R}^{1\times FT}$ of source i, i.e.,

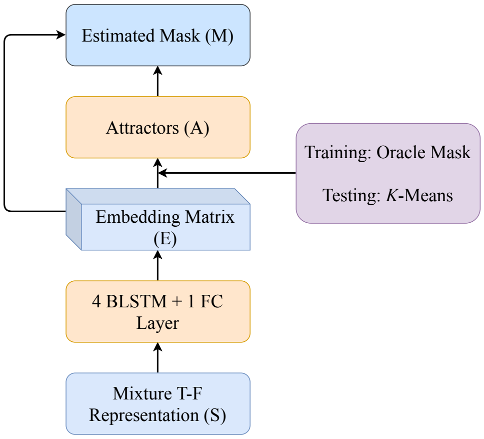

Fig. 2: DANet-based source separation. After [8].

$$a_i = \frac{m_i E^\top}{\sum_{f,t} m_i}. \tag{1}$$

The estimation of a mask is treated as a clustering problem. Firstly, compute the similarity of embedding of each T-F bin with the attractor of each source in embedding space. Then, estimate the mask by transforming this similarity to the probability distribution function (PDF). For source, i, estimated mask $\hat{m_i} \in \mathbb{R}^{1\times FT}$ is given by:

$$\hat{m_i} = Softmax(a_i E). \tag{2}$$

The loss function is a standard mean squared error (MSE), which directly optimizes the T-F representation of each source present in the mixture and it is given by [8]:

$$\mathcal{L}_{danet} = \frac{1}{P}\sum \|S \odot (m - \hat{m})\|_2^2. \tag{3}$$

B. Conventional CycleGAN

Let $x \in \mathbb{R}^N$, and $y \in \mathbb{R}^N$ be the cepstral features of source speaker (X), and target speaker (Y) speech, respectively, where N is the dimension of a feature vector. In CycleGAN, two generators are used: $G_{X\rightarrow Y}$, and $G_{Y\rightarrow X}$, where $G_{X\rightarrow Y}$ maps the cepstral features x to Y, whereas mapping $G_{Y\rightarrow X}$ does the opposite (i.e., y to X). In addition, we have two discriminators D_X and D_Y, whose role is to predict whether its input is from the distribution X, and Y or not, respectively. In CycleGAN, there are three types of losses, cycle-consistent

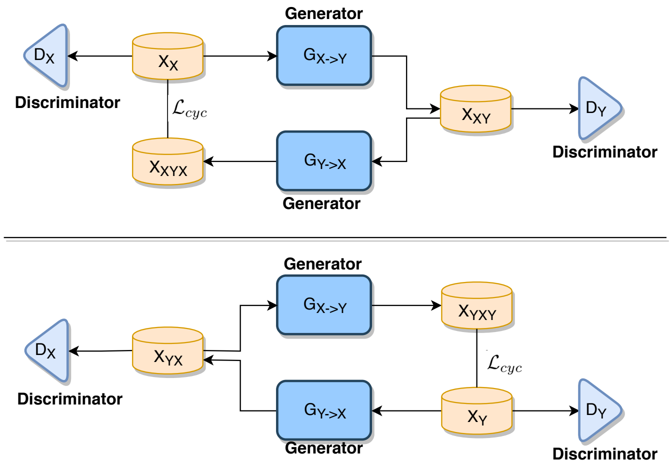

Fig. 3: Schematic representation of the CycleGAN-based singing style transfer system. After [14].

loss, adversarial loss, and identity loss, which is described next.

Adversarial loss: To make converted speech indistinguishable from the original target speech, we use adversarial loss. Here, we use least square error loss instead of traditional binary cross-entropy loss, which is defined as:

$$\begin{aligned}\mathcal{L}_{adv}(G_{X\to Y},D_Y) &= \mathbb{E}_{y\sim P_Y(y)}[(D_Y(y)-1)^2]\\ &+\mathbb{E}_{x\sim P_X(x)}[(D_Y(G_{X\to Y}(x)))^2].\end{aligned} \tag{4}$$

Cycle-consistent loss: The main idea behind this loss is to map the distribution between original and reconstructed data. In addition, this loss tries to preserve contextual information across different speech. This loss allows us to do non-parallel voice conversion. The loss is defined as:

$$\begin{aligned}&\mathcal{L}_{cyc}(G_{X\to Y},G_{Y\to X})\\ &=\mathbb{E}_{x\sim P_X(x)}[\|G_{Y\to X}(G_{X\to Y}(x))-x\|_1]\\ &+\mathbb{E}_{y\sim P_Y(y)}[\|G_{X\to Y}(G_{Y\to X}(y))-y\|_1].\end{aligned} \tag{5}$$

Identity-mapping loss: To encourage preservation of input linguistic content (as suggested in [23]), identity loss is used:

$$\begin{aligned}\mathcal{L}_{id}(G_{X\to Y},G_{Y\to X}) &= \mathbb{E}_{x\sim P_X(x)}[\|G_{Y\to X}(x)-x\|_1]\\ &+\mathbb{E}_{y\sim P_Y(y)}[\|G_{X\to Y}(y)-y\|_1].\end{aligned} \tag{6}$$

The total loss function is defined as:

$$\begin{aligned}&\mathcal{L}_{full}=\mathcal{L}_{adv}(G_{X\to Y},D_Y)+\mathcal{L}_{adv}(G_{Y\to X},D_X)\\ &+\lambda_{cyc}\mathcal{L}_{cyc}(G_{X\to Y},G_{Y\to X})+\lambda_{id}\mathcal{L}_{id}(G_{X\to Y},G_{Y\to X}),\end{aligned} \tag{7}$$

where the values of λ_{cyc}, and λ_{id} are 10 and 5, respectively.

C. Transfer Learning and Fine-Tuning of CycleGAN

Limited parallel data and synthesized data are a general problem in any machine learning architectures. To overcome these limitations, we have trained different CycleGAN models using transfer learning and fine-tuning based approach, as shown in Table I. For example, scenario 1, 2, and 4 are trained on the vocals of the MUSDB18 dataset (ground truth), vocals separated using DANet, vocals of the NUS48E dataset, respectively. Further, to take advantage of transfer learning, we fine-tuned our previously learned parameters of scenarios 1, 2, and 4. For example, in scenario 6, we pre-trained CycleGAN on vocals of the NUS48E dataset and fine-tuned with vocals of the MUSDB18 dataset (ground truth). For each scenario, we have trained 4 different models, in particular, for inter-gender (male-to-female, and female-to-male), and intra-gender (male-to-male, and female-to-female) speech conversion task for analysis.

TABLE I: Different system configuration based on transfer learning and fine-tuning

	Pretrained			Fine-tuned	
Systems	MUSDB18	NUS-48E	DANet	MUSDB18	DANet
Scenario 1	✓	-	-	-	-
Scenario 2	-	-	✓	-	-
Scenario 3	✓	-	-	-	✓
Scenario 4	-	✓	-	-	-
Scenario 5	-	✓	-	-	✓
Scenario 6	-	✓	-	✓	-
Scenario 7	-	✓	-	✓	✓

III. Experimental Results

A. Database and Feature Extraction

We make use of two databases for our experiments, namely, NUS-48E corpus, and MUSDB18. The NUS-48E corpus was first proposed in [22] and used to evaluate SINGAN architecture [2]. It consists of 48 English songs by 12 professional singers containing six female and six male singers. These songs include only vocals and have no additional instrumental music or noise. The MUSDB18 is composed of 150 tracks, of which 100 are used for training and 50 for testing [21]. The total duration of the dataset is 10 hours. The signals are stereophonic, encoded at 44.1 kHz, and are in a multi-track format comprising 5 stereo streams (drum, bass, other instruments, vocal, and the mixture of vocal and instruments). Since we do speaker-specific training, and the number of songs for one speaker in MUSDB18 is in the range of 1 to rarely 4, and for NUS, 4 songs for each speaker; makes us feel the need to augment the data. We split the song into smaller segments of $5s$ with an overlap of $1s$, also discarding the segments of complete noise or silence. The audio files are then converted to mono, 16-$bits$ per sample, and 16 kHz sampling frequency to extract cepstral, f_v and f_0 features using AHOCODER [24].

B. Architectural Details

1) DANet: We have used 3 bi-directional long short-term memory (Bi-LSTM) [25] layers, each with 600 hidden units with a dropout probability of 0.5. We set the embedding dimension to 20 as in [10], which results in 2580 hidden units in a fully-connected layer. ADAM algorithm is used for optimization with an initial learning rate of 0.001. The learning rate is halved if there is no decrease in validation loss for 3 epochs. To prevent a network from vanishing or exploiting gradient, the L$_2$ norm of gradient is clipped at 3. The model is trained for 80 epochs.

The Log-magnitude spectrogram is used as an input feature to the network. The input feature is computed using a short-time Fourier transform (STFT) with a window of 256 samples, 75 % overlap, and the square root of the Hanning window. It is split into 250 frames of non-overlapping segments, and feed as input to the network. Weiner filter-like mask (WFM) is used as an ideal mask [26], i.e.,

$$WFM_i = \frac{|s_i|^2}{\sum_{i=1}^{c} |s_i|^2}. \tag{8}$$

During testing, we have used K-means clustering algorithm to form an attractor for each source present in the speech mixture.

2) CycleGAN: Generators $G_{X \to Y}$, and $G_{Y \to X}$ follow the same configuration. In $G_{X \to Y}$, and $G_{Y \to X}$, contain 40, 512, and 40 neurons in the input layer, hidden layers, and output layer, respectively. All the layers are followed by Rectified Linear Unit (ReLU) activation function. All the discriminators follow the same configuration for both the architecture. D_X, and D_Y have 40, 512, and 1 neurons in the input layer, hidden layers, and output layer, respectively. Moreover, we use a batch size of 1000, as suggested in [27]. In all the discriminators, input layer, and all the hidden layers are followed by *ReLU* activation function, and output layer followed by *sigmoid* activation function.

C. Objective Evaluation

We have used three different objective measures to evaluate the efficiency of the DANet for music separation, namely, signal-to-distortion ratio (SDR), signal-to-artifacts ratio (SAR), and signal-to-interference ratio (SIR) [28]. Results are shown in Table II. Here, we are able to reproduce the results of the accomplishment described in the original paper [10]. However, results of vocals have been decreased as instead of 100 lengths of sequence while training, we have used 250 length sequence to increase the training speed (used in [10]). Therefore, results are not the best, however, this allows us to perform a more detailed analysis of CycleGAN fine-tuning as it's input features will have some amount of noise. Hence, we are able to measure the robustness of different CycleGAN systems.

TABLE II: Results of music separation for two-source condition

	SDR	SIR	SAR
Vocal	0.57	4.24	5.19
Accomplishment	9.03	21.83	9.45

TABLE III: Pitch analysis via NRMSE for different systems

Systems	FF	FM	MF	MM
Scenario 1	0.2255	**0.1943**	0.2528	0.2414
Scenario 2	0.2190	0.2052	0.2654	0.2362
Scenario 3	0.2090	0.2122	0.2556	0.2325
Scenario 4	**0.2019**	0.2047	**0.2261**	**0.1915**
Scenario 5	0.2194	0.2091	0.2644	0.2447
Scenario 6	0.2253	0.2009	0.2521	0.2155
Scenario 7	0.2150	0.2079	0.2628	0.2277

The note frequencies that the singer sings at and the voice converted singer sings at allows us to evaluate the singing quality of the singer [29]. To evaluate different CycleGAN systems objectively, we compute the Normalized Root Mean Squared Error (NRMSE) by the formula:

$$NRMSE = \frac{\sqrt{\frac{\sum_{i=1}^{n} |y_i - \hat{y_i}|^2}{n}}}{\bar{y}}, \tag{9}$$

where y_i indicates the note in i^{th} MIDI message of the original singer, $\hat{y_i}$ indicates the note in i^{th} MIDI message of the converted singer, n indicates the total length of notes in the note sequence, $\bar{y}$ indicates the maximum of mean note of the original singer, and the converted singer.

D. Subjective Evaluation

We have considered a total of 7 scenarios, as shown in Table I, and each scenario is comprised of a total of 4 models, i.e., male-male, female-female, and cross-gender conversion. To better determine the quality of results of our proposed models, we consider the 5-point Mean Opinion Score (MOS) [30]. We conducted two subjective tests on a total of 10 listeners (5 males and 5 females having age between 18 to 30 years, and having no known hearing impairment). In the first test, 5 $secs$ segments of 14 converted song were chosen in a way that two files were a result of each of our 7 scenarios to evaluate the naturalness of the converted song. In the second test, we randomly selected two files from each of our 28 models to evaluate speaker similarity of the converted song.

TABLE IV: MOS analysis for naturalness for different systems

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6	Scenario 7
MOS	3.68	2.55	2.82	**4.14**	3.45	2.73	2.64
STD	0.839	0.963	0.853	0.990	1.371	1.420	0.953

TABLE V: MOS analysis for speaker similarity for different systems

	FF		FM		MF		MM	
Systems	MOS	STD	MOS	STD	MOS	STD	MOS	STD
Scenario 1	2.70	1.13	**3.25**	1.16	2.90	1.07	2.35	1.04
Scenario 2	3.15	1.23	3.00	1.08	1.65	1.04	**2.80**	1.06
Scenario 3	2.75	1.16	2.75	0.79	**2.95**	0.83	**2.80**	0.83
Scenario 4	**3.60**	1.10	2.00	1.08	1.85	0.99	2.60	0.94
Scenario 5	3.30	0.92	3.05	1.19	2.65	1.18	2.55	0.69
Scenario 6	2.90	0.91	2.70	0.92	2.10	0.91	2.35	0.93
Scenario 7	3.10	1.37	2.90	1.21	2.50	1.10	2.70	1.08

From Table IV, it can be observed that scenario 4 outperforms every other scenario in terms of naturalness, and the same can also be observed in the objective analysis based on NRMSE. However, in terms of speaker similarity overall performance of scenario 1-3 is better than scenario 4. The reason behind such contradictory results could be due to the pre-training of scenario 4 is on NUS-48E, which only consists of normal (in terms of less deviations in pitch and rhythms) singing audios, however, pre-training of scenario 1 is on MUSDB18 dataset which is more complex. Ideally, scenario 5-7 should give better results in terms of speaker similarity as we have more speaker-specific data in NUS-48E compared to the MUSDB18, however, Table V shows the contradictory results, where scenario 2-3 performs relatively much better. The main reason behind this could be due to the *overfitting*. As we have very limited speaker-specific data in MUSDB18, testing data is made of the same speaker-pair as training data.

IV. Summary and Conclusions

In this paper, we proposed a transfer learning and fine-tuning based approach for non-parallel SVC in the presence of background music. We do an extensive analysis of seven different transfer learning scenarios using objective and subjective measures. We tested our model for the MUSDB18 dataset. To the best of authors' knowledge, the current methods have applied voice conversion for songs composed of vocals alone. For the first time in the literature, we introduce a transfer learning-based approach using GANs on a music dataset that contains background music and vocals. We achieve this by using a two-step approach of separating the vocals by DANet followed by voice conversion using CycleGAN along with introducing an additional novelty by fine-tuning using different permutations of NUS-48E, and MUSDB18 datasets. In the future, we plan to fully convert the song sung by the original speaker to the one sung by the target speaker, keeping background music, lyrics, and rhythm still intact (since our proposed model converts the speaker's identity, and maintains only the lyrics, and rhythm).

References

[1] K. Kobayashi, T. Toda, H. Doi, T. Nakano, M. Goto, G. Neubig, S. Sakti, and S. Nakamura, "Voice timbre control based on perceived age in singing voice conversion," *IEICE Transactions on Information and Systems*, vol. 97, no. 6, pp. 1419–1428, 2014.

[2] B. Sisman, K. Vijayan, M. Dong, and H. Li, "SINGAN: Singing voice conversion with generative adversarial networks," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Lanzhou, China, Nov 18-21, 2019, pp. 1–7.

[3] J. Sundberg and T. D. Rossing, "The science of singing voice," *The Journal of the Acoustical Society of America (JASA)*, vol. 87, no. 1, pp. 462–463, 1990.

[4] P. R. Cook, "Singing voice synthesis: History, current work, and future directions," *Computer Music Journal*, vol. 20, no. 3, pp. 38–46, 1996.

[5] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar 20-25, 2016, pp. 31–35.

[6] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. INTERSPEECH 2016*, San Francisco, CA, USA, Sep 8-12, 2016, pp. 545–549.

[7] D. Yu, M. Kolbæk, Z. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, NEW ORLEANS, LA, USA, Mar 5-9, 2017, pp. 241–245.

[8] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, April 2018.

[9] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, NEW ORLEANS, LA, USA, Mar 5-9, 2017, pp. 246–250.

[10] R. Kumar, Y. Luo, and N. Mesgarani, "Music source activity detection and separation using deep attractor network," in *Proc. INTERSPEECH 2018*, Hyderabad, India, Sep 2-6, 2018, pp. 347–351.

[11] J. chieh Chou, C. chieh Yeh, H. yi Lee, and L. shan Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," in *Proc. INTERSPEECH 2018*, Hyderabad, India, Sep 2-6, 2018, pp. 501–505.

[12] J. Zhang, Z. Ling, and L. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 540–552, 2020.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, Montreal, Canada, Dec 8-13, 2014, pp. 2672–2680.

[14] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv preprint arXiv:1711.11293*, 2017, {Last Accessed: January 28, 2020}.

[15] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved cyclegan-based non-parallel voice conversion," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 12-17, 2019, pp. 6820–6824.

[16] M. Patel, M. Parmar, S. Doshi, N. J. Shah, and H. A. Patil, "Adaptive generative adversarial network for voice conversion," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Lanzhou, China, Nov 18-21, 2019, pp. 1273–1281.

[17] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, Athens, Greece, Dec 18-21, 2018, pp. 266–273.

[18] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "StarGAN-VC2: Rethinking Conditional Methods for StarGAN-Based Voice Conversion," in *Proc. Interspeech 2019*, Graz, Austria, Sep 15-19, 2019, pp. 679–683.

[19] E. Nachmani and L. Wolf, "Unsupervised Singing Voice Conversion," in *Proc. Interspeech 2019*, Graz, Austria, Sep 15-19, 2019, pp. 2583–2587.

[20] X. Chen, W. Chu, J. Guo, and N. Xu, "Singing voice conversion with non-parallel data," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, San Jose, CA, USA, Aug 6-8, 2019, pp. 292–296.

[21] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017. [Online]. Available: https://sigsep.github.io/datasets/musdb.html, Last Accessed: January 28, 2020

[22] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, "The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Kaohsiung, Taiwan, Oct 29-Nov 1, 2013, pp. 1–9.

[23] 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 (ICCV)*, Venice, Italy, Oct 22-29, 2017, pp. 2223–2232.

[24] D. Erro, I. Sainz, E. Navas, and I. Hernáez, "Improved HNM-based vocoder for statistical synthesizers," in *Porc. INTERSPEECH 2011*, Florence, Italy, Aug 27-31, 2011, pp. 1809–1812.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, QLD, Australia, Apr 19-24, 2015, pp. 708–712.

[27] N. Shah, N. Shah, and H. Patil, "Effectiveness of generative adversarial network for non-audible murmur-to-whisper speech conversion," in *Proc. INTERSPEECH 2018*, Hyderabad, India, Sep 2-6, 2018, pp. 3157–3161.

[28] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[29] G. F. Welch, C. Rush, and D. Howard, "The singad (singing assessment and development) system: First applications in the classroom," *Proceedings of the Institute of Acoustics*, vol. 10, no. 2, pp. 179–185, 1988.

[30] I. ITU, "A method for subjective performance assessment of the quality of speech voice output devices," 1994, Available Online: {https://www.itu.int/rec/T-REC-P.85-199406-I/en}, {Last Accessed: January 28, 2020}.