Instrument-to-Instrument Timbre Conversion Using MaskCycleGAN

Wen-Hsing Lai
Department and Graduate Institute of Computer and Communication Engineering
National Kaohsiung University of Science and Technology
Kaohsiung City, Taiwan
lwh@nkust.edu.tw

Ching-Wei Hsieh
Department and Graduate Institute of Computer and Communication Engineering
National Kaohsiung University of Science and Technology
Kaohsiung City, Taiwan
f111110112@nkust.edu.tw

***Abstract*—This paper investigates instrument-to-instrument timbre conversion by employing the MaskCycleGAN architecture in conjunction with the MelGAN vocoder. The proposed framework enables the transformation of piano sounds into violin sounds using both parallel and non-parallel datasets. To examine the impact of training conditions, multiple model variants were implemented with different dataset types and residual channel settings. Objective evaluations, based on Mel-Cepstral Distortion (MCD) and Fréchet Audio Distance (FAD), demonstrate that models trained with parallel datasets achieve superior conversion performance. Subjective evaluations, including MOS and CMOS tests, further confirm the perceptual validity of the converted sounds. The results indicate that the proposed method reduces the technical barrier for performers by allowing seamless cross-instrument conversion and provides a promising tool for enhancing creativity in music production.**

Keywords—MaskCycleGAN, audio conversion, instrument-to-instrument, timbre conversion, MelGAN, piano to violin conversion

I. INTRODUCTION

With the continuous advancement of technology, audio conversion has become an increasingly prominent research area, such as speech emotion conversion [1] and singing voice conversion [2]. Audio conversion also demonstrates significant potential in music creation. Instrument-to-instrument timbre conversion has emerged as a new research direction, not only achieving the desired sound of target instruments but also fostering diversity and innovation in music creation. In music composition, the timbre of instruments is a crucial element, giving each work its unique character and opening up broader avenues for exploration. Different instruments represent distinct playing techniques and unique timbres. Composers aim to integrate these features to create novel and unique musical works. Through indepth exploration of instrument-to-instrument timbre conversion techniques, this research aspires to bring more inspiration and possibilities to music creation.

Compared to voice conversion, which has received significant attention and has been the subject of the Voice Conversion Challenge held in 2016, 2018, and 2020 [3]-[5], research on instrument-to-instrument timbre conversion is relatively recent. Voice conversion techniques have evolved considerably, beginning with approaches utilizing parallel data, such as spectral parameter trajectory maximum-likelihood estimation [6], spectral mapping using artificial neural networks [7], and Recurrent Temporal Restricted Boltzmann Machines (RTRBMs) [8]. Subsequently, methods for non-parallel data conversion were also proposed. Non-parallel data conversion methods are generally categorized into two main types. The first is feature disentanglement-based approaches, which aim to separate different factors of variation within a speech signal, such as linguistic content, speaker identity, and background noise, and by isolating these components, the model can independently manipulate specific attributes. For instance, L. Sun and H. Wang proposed a non-parallel voice conversion system that combines Phonetic PosteriorGrams (PPGs) obtained from an Automatic Speech Recognition (ASR) system with a Deep Bidirectional Long Short-Term Memory (DBLSTM) network [9]. The second category involves direct transformation-based methods, where the model learns to convert source to target directly. Notable examples include CycleGAN-VC [10], as well as its enhanced versions CycleGAN-VC2 [11], CycleGAN-VC3 [12], and MaskCycleGAN-VC [13].

Recently, instrument-to-instrument timbre conversion has started gaining attention as well. For example, the Differentiable Digital Signal Processing (DDSP) audio synthesis model [14] has been applied to instrument timbre conversion. By incorporating differentiable oscillators, filters, and reverberation components, DDSP enables high-quality audio synthesis with fewer data and parameters.

This study implements timbre conversion between instruments using the MaskCycleGAN architecture. MaskCycleGAN requires smaller datasets and results in a relatively compact model with faster processing times [13]. It incorporates the vocoder MelGAN [15] as a synthesizer for converting waveforms to Mel-spectrograms and vice versa. By adjusting various parameters and utilizing both parallel and non-parallel datasets for training, the experiments successfully realize audio conversion. Finally, subjective and objective evaluations are conducted to assess the outcomes.

The structure of this paper is as follows: Section 2 presents the method for instrument-to-instrument timbre conversion. Section 3 and 4 provide the experimental results and evaluation analysis. Concluding remarks are provided in the final section.

II. METHODOLOGY

The preprocessing procedure is follows: audio samples are first transformed into Mel-spectrograms via the Short-Time Fourier Transform (STFT). The mean and standard deviation of each Mel-spectrogram are computed for normalization. The resulting normalized values serve as the real input x to the MaskCycleGAN architecture.

Detailed descriptions of the MaskCycleGAN and MelGAN architectures are provided in the following subsections.

A. MaskCycleGAN

The concept of Generative Adversarial Networks (GANs) was introduced by Ian J. Goodfellow in 2014 [16]. The core idea involves two neural networks— a generator and a discriminator—trained in opposition. The generator aims to produce realistic fake data, while the discriminator attempts to distinguish between real and generated data. Building upon this concept, CycleGAN [17] was later proposed, and in 2018, Kaneko et al. introduced CycleGAN-VC [10], followed by the improved CycleGAN-VC2 in 2019 [11].

The CycleGAN-VC architecture incorporates three primary loss functions: adversarial loss, identity-mapping loss, and cycle consistency loss. The improved CycleGAN-VC2 model builds upon this structure by introducing a two-step adversarial loss, employing a 2-1-2D CNN architecture for the generator, and adopting a PatchGAN as the discriminator. In 2020, CycleGAN-VC3 [12] was introduced to address a key limitation of previous CycleGAN-based voice conversion models—namely, the inability to fully preserve the time-frequency structure of mel-spectrograms during conversion. To overcome this limitation, CycleGAN-VC3 draws inspiration from semantic image synthesis and adopts the SPatially-Adaptive (DE)normalization (SPADE) technique [18]. It further proposes Time-Frequency Adaptive Normalization (TFAN), which utilizes convolutional neural networks (CNNs) to learn the time-frequency structure of the source mel-spectrogram. This information is then used to modulate the scale and bias of the transformation features, thereby preserving the source structure in the converted mel-spectrogram. In 2021, MaskCycleGAN-VC [13] was proposed as a further extension. While still operating on mel-spectrogram representations, it introduces a novel strategy called "Filling-In Frames" (FIF), which guides the generator to recover missing frames by leveraging surrounding frame information.

In the MaskCycleGAN framework, the input mel-spectrogram is denoted as x, and a mask m of the same dimensions is applied through element-wise multiplication to obtain the masked spectrogram $\hat{x}$, as shown in Eq. (1).

$$\hat{x} = x \cdot m \tag{1}$$

The masked spectrogram $\hat{x}$, together with the mask m, is channel-wise concatenated (denoted by $concat$) and passed to the forward generator $G_{X\to Y}^{mask}$. Conditioned on the mask m, $G_{X\to Y}^{mask}$ can selectively infer and reconstruct the absent regions in the input spectrogram. The conditionally informed generator then outputs the converted mel-spectrogram y', as in Eq. (2).

$$y' = G_{X\to Y}^{mask}(concat(\hat{x}, m)) \tag{2}$$

Due to the lack of parallel data for direct supervision, the converted result cannot be directly compared with a ground-truth target. Therefore, a reverse generator $G_{Y\to X}^{mask}$ is used in a cycle-consistent fashion to reconstruct x'', as defined in Eq. (3), where m' is assumed to be a matrix of ones, indicating fully filled frames.

$$x'' = G_{Y\to X}^{mask}(concat(y', m')) \tag{3}$$

Subsequently, a cycle consistency loss is applied between the original mel-spectrogram x and the reconstructed spectrogram x''. Additionally, a second adversarial loss is imposed on x'', as defined in Eq. (4).

$$L_{mcyc}^{X\to Y\to X} = \mathbb{E}_{x\sim P_X, m\sim P_M}[\| x'' - x \|_1] \tag{4}$$

B. MelGAN

Following the conversion process performed by MaskCycleGAN, MelGAN is employed as the vocoder to synthesize the converted speech. In MelGAN [15], the mel-spectrogram serves as the primary representation for conversion. By optimizing the parameters of both the generator and the discriminator, the model is trained to reconstruct the original waveform within the GAN framework.

As illustrated in Fig. 1, the architecture of the MelGAN generator takes the mel-spectrogram as input, which is first processed by a convolutional layer, followed by a series of upsampling layers and residual stacks within the generator. The output from these modules is then passed through a final convolutional layer to produce the synthesized audio waveform. The total upsampling factor of the generator is calculated as 10×10×2×2×2=800.

Fig. 2 illustrates the MelGAN discriminator architecture, which employs three discriminators operating at different temporal resolutions. The original waveform is fed to the first discriminator, while downsampled versions—obtained via average pooling—are provided to the others. Each discriminator comprises a stack of one-dimensional convolutional layers, beginning with a large-kernel convolution to capture broad temporal context. This is followed by progressively smaller-kernel convolutional blocks with a fixed stride 2 to reduce temporal resolution and expand the receptive field. Non-linear activations LeakyReLU are applied to all layers, and weight normalization is used to stabilize adversarial training. Each discriminator produces a local real/fake decision map. The outputs across scales are aggregated to compute the discriminator loss in adversarial learning.

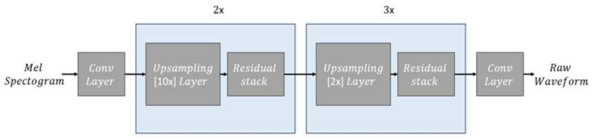


Fig. 1. Schematic illustration of the MelGAN generator architecture

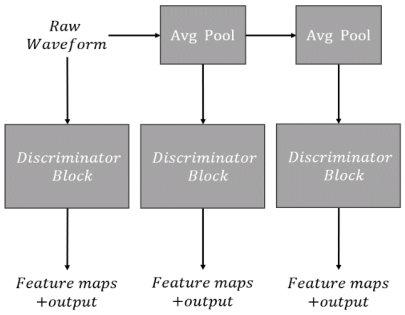


Fig. 2. Schematic illustration of the MelGAN discriminator architecture

III. EXPERIMENTS

In this study, the experimental setup focuses on musical instrument sound conversion from piano to violin. The MaskCycleGAN architecture is employed to perform the instrument voice conversion between piano and violin, while MelGAN is adopted as the vocoder for waveform synthesis.

A. Datasets

The dataset includes both parallel and non-parallel types. Since parallel datasets are relatively difficult to obtain, they are generated through MIDI-based conversions. The parallel dataset consists of 48 piano and 48 violin recordings with relatively simple melodies, among which 5 pieces are reserved as the test set. The non-parallel dataset contains 195 piano and 195 violin pieces with more complex melodic structures. The piano and violin data in the non-parallel set come from different sources: the piano recordings are taken from the MAESTRO (MIDI and Audio Edited for Synchronous TRacks and Organization) dataset [19], which comprises approximately 200 hours of piano performances and corresponding MIDI files. The violin recordings are from the CocoChorales dataset [20], generated by Yusong Wu, Josh Gardner, and others using the Chamber Ensemble Generator. All datasets were resampled to 16 kHz. Table I summarizes the characteristics of the datasets.

B. Training of MaskCycleGAN

During the training of MaskCycleGAN, the audio sampling rate was set to 16 kHz, the mask size was 25, the number of mel-spectrogram channels was 64, and the number of training iterations was set to 5000. To evaluate the impact of different training settings on the quality of the generated audio, four training versions were created using combinations of parallel or non-parallel datasets and residual channels set to either 256 or 512, as summarized in Table II.

The training loss dynamics of the MaskCycleGAN architecture are analyzed below. The horizontal axis represents training time in hours. As shown in Figs. 3 and 4, the loss curves for MCGnp512 and MCGp512 illustrate the evolution of various losses during training, including (a) generator loss, (b) generator X→Y loss and generator Y→X loss, (c) cycle consistency loss and mapping loss, and (d) discriminator loss, discriminator D_x loss, and discriminator D_y loss.

TABLE I. DATASETS

Dataset Type	Parallel		Non-Parallel	
Instrument	Piano	Violin	Piano	Violin
Sampling Rate	16kHZ	16kHZ	16kHZ	16kHZ
Number of Pieces	48	48	195	195
Duration Range	9s~278s	9s~278s	241s~7689s	13s~39s
Source	MIDI	MIDI	MAESTRO	CocoChorlaes

TABLE II. VERSIONS OF MASKCYCLEGAN

Version	Dataset	Residual Channels
MCGnp256	nonparallel	256
MCGnp512	nonparallel	512
MCGp256	parallel	256
MCGp512	parallel	512

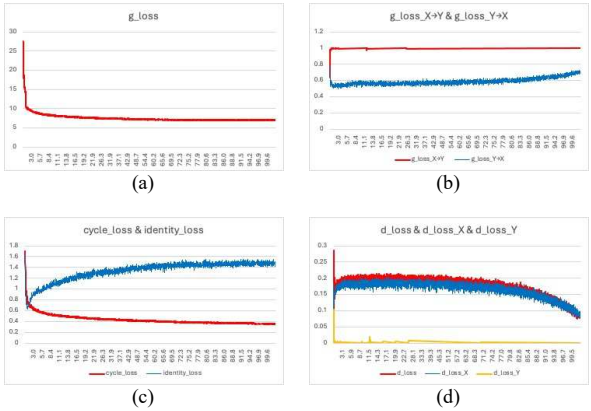


Fig. 3. MCGnp512 Loss Plots: (a) Generator, (b) Generator X→Y and Y→X, (c) Cycle-consistency and Mapping, (d) Discriminator, D_x and D_y

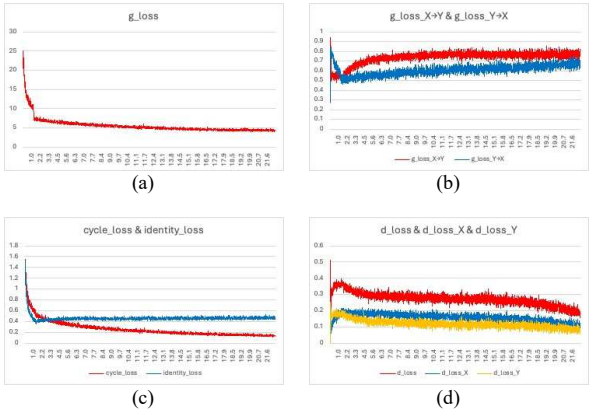


Fig. 4. MCGp512 Loss Plots: (a) Generator, (b) Generator X→Y and Y→X, (c) Cycle-consistency and Mapping, (d) Discriminator, D_x and D_y

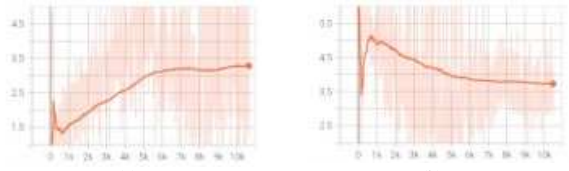
TABLE III. TRAINING PARAMETERS OF MELGAN VOCODER

Parameter	Value
Training Iterations	60000
Sampling Rate	16kHz
Mel Spectrogram Channels	64
Segment Length	8000
Batch Size	64
FFT Size	1024
Window Length	1024
Upsampling Layers	800
Upsampling Ratios	$10 \times 10 \times 2 \times 2 \times 2$

C. Training of MelGAN

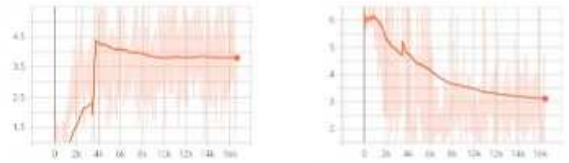
To suit the voice conversion task, the sampling rate is set to 16 kHz, and the mel-spectrogram channel size and audio duration are adjusted accordingly. MelGAN is trained based on these settings, as summarized in Table III.

Using the training parameters specified in Table III, MelGAN was trained with both parallel and non-parallel conversion datasets. The corresponding generator and discriminator losses were measured with respect to the number of iterations, as illustrated in Figs. 5 and 6.



(a) (b)

Fig. 5. (a) Generator loss and (b) discriminator loss plots for the parallel conversion dataset

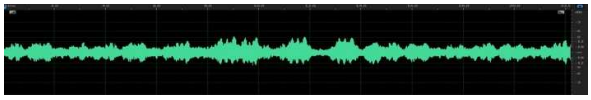


(a) (b)

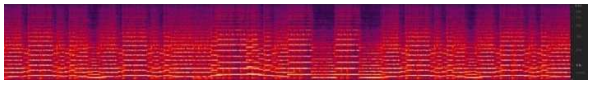
Fig. 6. (a) Generator loss and (b) discriminator loss plots for the non-parallel conversion dataset

TABLE IV. VERSIONS OF THE INSTRUMENT SOUND CONVERSION MODEL

Version	MaskCycleGAN	Residual Channels	MelGAN
MCGnp256_MGnp	nonparallel	256	nonparallel
MCGnp256_MGp	nonparallel	256	parallel
MCGnp512_MGnp	nonparallel	512	nonparallel
MCGnp512_MGp	nonparallel	512	parallel
MCGp256_MGnp	parallel	256	nonparallel
MCGp256_MGp	parallel	256	parallel
MCGp512_MGnp	parallel	512	nonparallel
MCGp512_MGp	parallel	512	parallel



(a)



(b)

Fig. 7. (a) Waveform and (b) Mel-spectrogram of MCGp512_MGp

D. Versions of the Instrument Sound Conversion Model

Instrument-to-instrument sound conversion between piano and violin is conducted using the MaskCycleGAN architecture, with MelGAN employed as the vocoder. By combining either parallel or non-parallel datasets for training both MaskCycleGAN and MelGAN, and configuring the residual channels to either 256 or 512, eight different instrument sound conversion versions are generated, as summarized in Table IV. Fig. 7 presents an example of the converted instrument waveform along with the corresponding mel-spectrogram obtained from MCGP512_MGP.

IV. EVALUATION

Human perception of sound varies greatly among individuals, rendering auditory experience inherently subjective. Consequently, this study employs both subjective and objective evaluations as the criteria for assessing the experimental outcomes.

A. Objective Evaluation Metrics

The objective evaluation in this study adopts two widely used metrics: Mel-Cepstral Distortion (MCD) [21] and Fréchet Audio Distance (FAD) [22].

MCD, a common evaluation metric in voice conversion and speech synthesis, measures the difference between synthesized speech and target speech. It is particularly effective in reflecting perceived differences in audio quality. In this experiment, MCD is computed between audio converted from piano to violin and the corresponding target violin recordings; lower MCD values indicate higher similarity between the converted and target audio.

FAD is an objective metric for evaluating the perceptual quality of audio signals, introduced by Google Research as an audio-domain counterpart to the Fréchet Inception Distance (FID) [23] used in image quality assessment. Unlike traditional distortion-based measures, FAD does not require the original (clean) audio as a reference. Instead, it computes the statistical distance between feature distributions extracted from the audio under evaluation and a set of high-quality reference recordings. The features are derived from a pretrained VGGish network, which captures high-level semantic and perceptual characteristics of the audio. The distance is calculated using the Fréchet distance between multivariate Gaussian distributions fitted to the feature sets. Lower FAD values indicate closer alignment of the evaluated audio with the reference domain, thus implying better perceptual quality.

B. Subjective Evaluation Metrics

For the subjective evaluation, two perceptual tests were conducted: the Mean Opinion Score (MOS) and the Comparison Mean Opinion Score (CMOS) test. Both tests were administered via online questionnaires distributed to participants, who were instructed to provide ratings according to the specified evaluation criteria.

In the MOS test, participants were asked to evaluate the perceptual quality of the presented audio samples on a five-point scale, where a score of 1 indicates "very poor" quality and a score of 5 indicates "excellent" quality. Audio samples generated in the experiments were embedded directly into the questionnaire, and ratings were collected through an online survey form. The detailed interpretation of the rating scale is provided in Table V.

According to Annex E of the ITU-T Recommendation P.800 [24], the Comparative Mean Opinion Score (CMOS) method serves as an alternative subjective metric for quality assessment. In a CMOS test, listeners compare two different versions of the test audio and assign a score based on their perceived preference. The original scoring scale ranges from –3 to +3, corresponding to seven discrete levels. To reduce the cognitive load on participants when discerning subtle differences and assigning scores, this study simplifies the scale to three levels: –1 (worse), 0 (equal), and +1 (better).

C. Objective Evaluation Results

In this experiment, objective evaluation was conducted using two metrics: MCD and FAD. The scores of the generated audio for each version of the test set are presented in Table VI.

Through this objective evaluation, it was found that the scores of MCGp256_MGp and MCGp512_MGp were consistently lower than those of other experimental methods, indicating superior performance. Therefore, the approaches using parallel conversion datasets in both MaskCycleGAN and MelGAN outperform those employing non-parallel conversion datasets.

D. Subjective Evaluation Results

In this experiment, a subjective evaluation was conducted using the MOS metric to assess the conversion of five piano pieces (A to E) of the test set of the parallel dataset into violin sounds. The conversions were performed by two models: MCGp256_MGp and MCGp512_MGp. Violin sounds synthesized directly from the MIDI files of the test set were included as a reference baseline for comparison. Fifteen participants rated the audio samples based on their subjective perception. The MOS results for the model-generated conversions by MCGp256_MGp and MCGp512_MGp are summarized in Tables VII (a) and (b), while the MOS scores for the MIDI-synthesized violin sounds are presented in Table VII (c).

Based on the MOS evaluation results presented in Tables VII (a), (b), and (c), the average scores were calculated as follows: 4.21 for MCGp256_MGp, 4.19 for MCGp512_MGp, and 4.51 for the MIDI-generated violin sounds.

Furthermore, a CMOS evaluation was conducted to compare the MCGp512_MGp and MIDI-generated violin sounds for the five test-set pieces. Fifteen participants were asked to select the version they perceived as superior. The results, shown in Table VIII, indicate that 74.67% of the participants preferred the MIDI-generated violin sounds, while 25.33% favored the MCGp512_MGp outputs.

V. Conclusions and Future Works

The objective of this study is to enable effortless conversion from one musical instrument sound to another by employing the MaskCycleGAN network architecture in conjunction with the MelGAN vocoder. To this end, experiments were conducted using both parallel and non-parallel piano-to-violin datasets, and the results were compared to evaluate the influence of dataset type on the conversion performance.

In the present experiments, superior results were achieved for monophonic instrument sound conversion. However, musical performances often involve complex melodies rather than such simplified cases. For more intricate melodies, the conversion quality is significantly reduced. Addressing this limitation will be the focus of our future work.

TABLE V. MOS Rating Criteria for Violin Timbre Quality

MOS	Grade	Violin Timbre Quality Description
5	Excellent	Timbre remains stable with good rhythmic consistency and no audible distortion.
4	Good	Timbre is stable with good rhythmic consistency; distortion is negligible.
3	Fair	Timbre quality is slightly degraded; rhythm is unstable, and distortion is noticeable.
2	Poor	Timbre quality is severely degraded; rhythm is unstable, and distortion is pronounced.
1	Very Poor	Timbre and rhythm are unrecognizable; distortion is excessively severe.

TABLE VI. Results for MIDI-to-Violin Conversion

Version	MCD	FAD
MCGnp256_MGnp	21.53	9.74
MCGnp256_MGp	19.48	9.79
MCGnp512_MGnp	25.90	10.68
MCGnp512_MGp	22.49	9.19
MCGp256_MGnp	16.41	5.31
MCGp256_MGp	4.98	**0.70**
MCGp512_MGnp	15.37	5.12
MCGp512_MGp	**4.81**	0.71

TABLE VII. MOS evaluation vote counts for (a) MCGp256_MGp (b) MCGp512_MGp (c) MIDI-synthesized violin

(a)

Audio\MOS	1	2	3	4	5
A	0	0	3	5	7
B	0	0	4	4	7
C	0	0	0	5	10
D	0	0	4	4	7
E	1	0	4	7	3
Percentage	1.33%	0%	20%	33.33%	45.33%

(b)

Audio\MOS	1	2	3	4	5
A	0	1	1	5	8
B	0	0	1	4	10
C	0	0	3	6	6
D	0	1	3	4	7
E	2	0	2	8	3
Percentage	2.66%	2.66%	13.33%	36%	45.33%

(c)

Audio\MOS	1	2	3	4	5
A	0	0	1	2	12
B	0	0	0	5	10
C	0	0	0	5	10
D	0	1	1	6	7
E	0	2	1	4	8
Percentage	0%	4%	4%	29.33%	62.66%

TABLE VIII. CMOS evaluation vote counts for MCGp512_MGp vs. MIDI-generated violin sounds

Audio	MCGp512_MGp	MIDI generated
A	3	12
B	3	12
C	5	10
D	5	10
E	3	12
Percentage	25.33%	74.67%

Acknowledgment

Part of this work was supported by the Ministry of Science and Technology, Taiwan under Contract MOST 111-2221-E-992-075.

References

[1] Carl Robinson, Nicolas Obin, and Axel Roebel, "Sequence-to-sequence Modelling of F0 for Speech Emotion Conversion," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Pro-cessing (ICASSP), 12-17 May 2019, pp. 6830-6834, doi: 10.1109/ICASSP.2019.8683865.

[2] Koki Senda, Yukiya Hono, Kei Sawada, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda, "Singing Voice Conversion Using Posted Waveform Data on Music Social Media," in 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 12-15 Nov. 2018 2018, pp. 1913-1917, doi: 10.23919/APSIPA.2018.8659568.

[3] Tomoki Toda, Ling-Hui Chen, Daisuke Saito, Fernando Villavicencio, Mirjam Wester, Zhizheng Wu, Junichi Yamagishi, "The Voice Conversion Challenge 2016," INTERSPEECH 2016, September 8–12, 2016, San Francisco, USA, pp. 1632-1636.

[4] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, Zhenhua Ling, "The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods," The Speaker and Language Recognition Workshop (Odyssey 2018), 26-29 June 2018, Les Sables d'Olonne, France, pp. 195-202.

[5] Zhao Yi, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhenhua Ling, Tomoki Toda, "Voice Conversion Challenge 2020 – Intra-lingual semi-parallel and cross-lingual voice conversion –," Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020, 30 October 2020, Shanghai, China, pp. 80-98.

[6] 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, November 2007, doi: 10.1109/TASL.2007.907344.

[7] Srinivas Desai, Alan W. Black, B. Yegnanarayana, and Kishore Prahallad, "Spectral Mapping Using Artificial Neural Networks for Voice Conversion," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 5, pp. 954-964, July 2010, doi: 10.1109/TASL.2010.2047683.

[8] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki, "Voice Conversion Using RNN Pre-Trained by Recurrent Temporal Restricted Boltzmann Machines," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 3, pp. 580-587, 2015, doi: 10.1109/TASLP.2014.2379589.

[9] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," 2016 IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, USA, 11-15 July 2016, pp. 1-6, DOI: 10.1109/ICME.2016.7552917.

[10] Takuhiro Kaneko and Hirokazu Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in Proc. EUSIPCO, 2018, pp. 2100–2104.

[11] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "CycleGan-VC2: Improved CycleGan-based Non-parallel Voice Conversion," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 12-17 May 2019 2019, pp. 6820-6824, doi: 10.1109/ICASSP.2019.8682897.

[12] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "CycleGAN-VC3: Examining and Improving CycleGAN-VCs for Mel-spectrogram Conversion," INTERSPEECH 2020, October 25–29, 2020, Shanghai, China, pp. 2017-2021.

[13] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, Nobukatsu Hojo, "Maskcyclegan-VC: Learning Non-Parallel Voice Conversion with Filling in Frames," in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6-11 June 2021 2021, pp. 5919-5923, doi: 10.1109/ICASSP39728.2021.9414851.

[14] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts, "DDSP: Differentiable digital signal processing," ICLR 2020.

[15] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, Aaron Courville, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," Proceedings of the 33rd International Conference on Neural Information Processing Systems, December 2019, Article No.: 1335, Pages 14910 - 14921.

[16] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozairy, Aaron Courville, Yoshua Bengioz, "Generative Adversarial Networks," June 2014, Advances in Neural Information Processing Systems 3(11), DOI:10.1145/3422622.

[17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in 2017 IEEE International Conference on Computer Vision (ICCV), 22-29 Oct. 2017, pp. 2242-2251, doi: 10.1109/ICCV.2017.244.

[18] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu, "Semantic Image Synthesis with Spatially-Adaptive Normalization," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15-20 June 2019, Long Beach, CA, USA, pp. 2332-2341.

[19] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dielemany, Erich Elsen, Jesse Engel, and Douglas Eck, "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset," ICLR 2019.

[20] YusongWu, Josh Gardner, Ethan Manilow, Ian Simon, Curtis Hawthorne, Jesse Engel, "The Chamber Ensemble Generator: Limitless High-Quality MIR Data via Generative Modeling," arXiv preprint arXiv:2209.14458.

[21] Robert F. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, 19-21 May 1993, Victoria, BC, Canada, pp. 125-128.

[22] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, Matthew Shari, "Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms," INTERSPEECH 2019, September 15–19, 2019, Graz, Austria, pp. 2350-2354.

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Sepp Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," Neural Information Processing Systems (NIPS), Volume: 30, December 2017, Long Beach, California.

[24] P.800: Methods for Subjective Determination of Transmission Quality. https://www.itu.int/rec/T-REC-P.800-199608-I. Accessed Jan. 09, 2021.