한국음향학회지 제41권 제3호 pp. 351~358 (2022)
The Journal of the Acoustical Society of Korea Vol.41, No.3 (2022)
https://doi.org/10.7776/ASK.2022.41.3.351

pISSN : 1225-4428
eISSN : 2287-3775

다수 화자 한국어 음성 변환 실험

Many-to-many voice conversion experiments using a Korean speech corpus

육동석,[1†] 서형진,[1] 고봉구,[1] 유인철[1]

(Dongsuk Yook,[1†] HyungJin Seo,[1] Bonggu Ko,[1] and In-Chul Yoo[1])

[1]고려대학교 컴퓨터학과 인공지능연구실

(Received March 16, 2022; revised April 29, 2022; accepted May 13, 2022)

초 록: 심층 생성 모델의 일종인 Generative Adversarial Network(GAN)과 Variational AutoEncoder(VAE)는 비병렬 학습 데이터를 사용한 음성 변환에 새로운 방법론을 제시하고 있다. 특히, Conditional Cycle-Consistent Generative Adversarial Network(CC-GAN)과 Cycle-Consistent Variational AutoEncoder(CycleVAE)는 다수 화자 사이의 음성 변환에 우수한 성능을 보이고 있다. 그러나, CC-GAN과 CycleVAE는 비교적 적은 수의 화자를 대상으로 연구가 진행되어왔다. 본 논문에서는 100 명의 한국어 화자 데이터를 사용하여 CC-GAN과 CycleVAE의 음성 변환 성능과 확장 가능성을 실험적으로 분석하였다. 실험 결과 소규모 화자의 경우 CC-GAN이 Mel-Cepstral Distortion(MCD) 기준으로 4.5 % 우수한 성능을 보이지만 대규모 화자의 경우 CycleVAE가 제한된 학습 시간 안에 12.7 % 우수한 성능을 보였다.

핵심용어: 음성 변환, Conditional Cycle-consistent Generative Adversarial Network (CC-GAN), Cycle-Consistent Variational AutoEncoder (CycleVAE), Generative Adversarial Network (GAN), Variational AutoEncoder (VAE)

ABSTRACT: Recently, Generative Adversarial Networks (GAN) and Variational AutoEncoders (VAE) have been applied to voice conversion that can make use of non-parallel training data. Especially, Conditional Cycle-Consistent Generative Adversarial Networks (CC-GAN) and Cycle-Consistent Variational AutoEncoders (CycleVAE) show promising results in many-to-many voice conversion among multiple speakers. However, the number of speakers has been relatively small in the conventional voice conversion studies using the CC-GANs and the CycleVAEs. In this paper, we extend the number of speakers to 100, and analyze the performances of the many-to-many voice conversion methods experimentally. It has been found through the experiments that the CC-GAN shows 4.5 % less Mel-Cepstral Distortion (MCD) for a small number of speakers, whereas the CycleVAE shows 12.7 % less MCD in a limited training time for a large number of speakers.

Keywords: Voice conversion, Conditional Cycle-Consistent Generative Adversarial Network (CC-GAN), Cycle-Consistent Variational AutoEncoder (CycleVAE), Generative Adversarial Network (GAN), Variational AutoEncoder (VAE)

PACS numbers: 43.72.Ja, 43.72.Bs

I. 서 론

음성 변환은 어떤 내용을 말했는지에 관한 언어 정보를 유지한 채 원본 화자의 목소리를 목표 화자의 목소리로 변환하는 것을 의미한다. 음성 변환 기술은 언어 장애인의 발음 보조, 발음 변환을 통한 외

†**Corresponding author:** Dongsuk Yook (yook@korea.ac.kr)
Artificial Intelligence Laboratory, Department of Computer Science and Engineering, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Republic of Korea
(Tel: 82-2-3290-3202)

"이 논문은 2019년도 한국음향학회 음성통신 및 신호처리 학술대회와 2021년도 한국음향학회 추계학술발표대회에서 발표한 내용을 확장한 것임.[1,2]"

국어 교육, 음성 합성 성능 향상 등의 응용 분야에 활용될 수 있다. 또한, 최근 많이 사용되는 음성 기반 인터페이스 기기에서 사용자의 목소리를 변환시켜 익명화함으로써 개인 정보 보호에 활용될 수도 있다.[3]

최근에는 심층 신경망(Deep Neural Network, DNN)을 기반으로 한 음성 변환 방법이 활발히 연구되고 있다. 특히, Generative Adversarial Network(GAN)[4]이나 Variational AutoEncoder(VAE)[5] 기반의 음성 변환 방법은 원본 화자와 목표 화자가 동일한 내용을 발화한 병렬 학습 데이터를 필요로 하지 않기 때문에 많은 학습 데이터를 활용할 수 있다는 장점이 있다. GAN 기반의 음성 변환에서는 두 개의 GAN을 연결하여 원본 화자 음성 데이터를 목표 화자 음성 데이터로 변환한 후 이것을 다시 원본 화자 음성 데이터로 복원할 수 있도록 학습하여 언어 정보를 유지한 채 목소리만 변환할 수 있게 하는 Cycle-Consistent Adversarial Network(CycleGAN)[6] 기반의 음성 변환이 우수한 성능을 보여주고 있다.[7-13] VAE 기반 음성 변환에서는 목표 화자 정보를 디코더의 추가 입력으로 활용하여 다수 화자 사이의 음성 변환 학습에 효율성을 보여주고 있다.[14-19]

CycleGAN 기반의 음성 변환은 변환된 음성의 음질이 상대적으로 높은 반면 다수 화자 사이의 음성 변환에 오랜 학습 시간이 필요하다는 단점이 있다. VAE 기반의 음성 변환은 학습 시간이 짧은 반면 변환된 음성의 음질이 상대적으로 낮다는 단점이 있다. 본 논문에서는 CycleGAN 기반 음성 변환 중에서 다수 화자 사이의 음성 변환에 비교적 효율적인 Conditional Cycle-Consistent Generative Adversarial Network(CC-GAN)과,[10,11] VAE 기반 음성 변환 중에서 명시적 변환 경로 학습을 추가하여 변환된 음성의 음질을 개선한 Cycle-Consistent Variational AutoEncoder(CycleVAE)를[19] 한국어 음성 변환 실험을 통하여 비교한다. 두 방법 모두 기존에는 비교적 적은 수의 화자를 대상으로 연구가 진행되어왔으나, 본 논문에서는 100명의 한국어 화자 데이터를 사용하여 CC-GAN과 CycleVAE의 음성 변환 성능과 확장 가능성을 실험적으로 분석한다.[20] 특히, 화자 수와 학습 시간에 따른 변환 음성 품질을 실험적으로 비교하여 CC-GAN과 CycleVAE의 효율성과 성능을 분석한다.

본 논문의 구조는 다음과 같다. II장에서는 심층 신경망 기반의 심층 생성 모델을 설명하고, CC-GAN과 CycleVAE를 이용한 다수 화자 음성 변환을 설명한다. III장에서는 앞서 언급한 음성 변환 방법들을 다수 화자 한국어 데이터를 이용하여 실험하고, IV장에서 결론을 맺는다.

II. 다수 화자 음성 변환 모델

II장에서는 다수 화자 음성 변환 실험에 사용될 CC-GAN과 CycleVAE를 설명한다.

2.1 CC-GAN

GAN은 생성자와 구분자로 불리는 두 신경망이 서로 경쟁하는 방식으로 학습하는 생성 모델이다. 학습에 성공하면 생성자는 구분자에 의해서 구분되기 어려운 실제 데이터와 유사한 데이터를 생성할 수 있게 된다. GAN의 학습에는 다음과 같은 목적 함수를 사용한다.

$$L_{\mathrm{GAN}}(G,D;x,y)=\log D(x)+\log(1-D(G(y))), \tag{1}$$

여기서 G는 생성자, D는 구분자, x는 실제 데이터, y는 임의의 노이즈를 의미한다. 음성 변환의 경우 x는 목표 화자의 음성이고, y는 원본 화자의 음성이 된다.

CycleGAN은 두 개의 GAN으로 입력 데이터를 변형한 후 다시 원래 입력 데이터를 복원할 수 있도록 학습한다. 이를 위해서 GAN 목적 함수 이외에 추가적으로 다음과 같은 cycle consistency loss를 사용한다.

$$L_{\mathrm{cycle}}(G_X,G_Y;x,y)=\|G_X(G_Y(x))-x\|_1+\|G_Y(G_X(y))-y\|_1, \tag{2}$$

여기서 G_X와 G_Y는 각각 화자 X와 화자 Y의 음성 데이터로 변환하는 생성자, x는 화자 X의 음성 데이터, y는 화자 Y의 음성 데이터, $\|\cdot\|_1$은 L1 norm을 의미한다. 화자 X와 Y는 화자 식별 벡터로 표현되는데, 일반적으로 one-hot 벡터를 사용한다.

G_X에 화자 X의 음성 데이터 x가 입력된 경우와 G_Y에 화자 Y의 음성 데이터 y가 입력된 경우 각 생성자는 입력 데이터를 변환없이 그대로 출력해야 한다. 이를 위해서 다음과 같은 identity mapping loss를 추가적으로 사용한다.

$$L_{\text{identity}}(G_X, G_Y; x, y) = \| G_X(x) - x \|_1 + \| G_Y(y) - y \|_1. \tag{3}$$

화자 X의 음성 데이터 x와 화자 Y의 음성 데이터 y가 주어졌을 때, CycleGAN의 최종 목적 함수는 다음과 같다.

$$\begin{aligned} L_{\text{CycleGAN}}(G_X, D_X, G_Y, D_Y; x, y) = \\ L_{\text{GAN}}(G_X, D_X; x, y) + L_{\text{GAN}}(G_Y, D_Y; y, x) + \\ \lambda_1 L_{\text{cycle}}(G_X, G_Y; x, y) + \\ \lambda_2 L_{\text{identity}}(G_X, G_Y; x, y), \end{aligned} \tag{4}$$

여기서 D_X와 D_Y는 각각 화자 X와 화자 Y의 음성 데이터를 판별하는 구분자, λ_1과 λ_2는 L_{cycle}과 L_{identity}의 상대적인 가중치이다.

CycleGAN은 두 화자 사이의 음성 변환에서 우수한 성능을 보이는데, 여기에 원본 화자와 목표 화자 정보를 조건으로 제공하여 다수 화자 음성 변환 모델로 확장한 것이 CC-GAN이다. 화자 X의 음성 데이터 x와 화자 Y의 음성 데이터 y가 주어졌을 때, CC-GAN의 목적 함수는 다음과 같다.

$$\begin{aligned} L_{\text{CC-GAN}}(G, D; x, y, X, Y) = \\ L'_{\text{GAN}}(G, D; x, y, X, Y) + \\ L'_{\text{GAN}}(G, D; y, x, Y, X) + \\ \lambda_1 L'_{\text{cycle}}(G; x, y, X, Y) + \\ \lambda_2 L'_{\text{identity}}(G; x, y, X, Y), \end{aligned} \tag{5}$$

$$L'_{\text{GAN}}(G, D; x, y, X, Y) = \log D(x, X) + \log(1 - D(G(y, Y, X), X)), \tag{6}$$

$$\begin{aligned} L'_{\text{cycle}}(G; x, y, X, Y) = \\ \| G(G(x, X, Y), Y, X) - x \|_1 + \\ \| G(G(y, Y, X), X, Y) - y \|_1, \end{aligned} \tag{7}$$

$$\begin{aligned} L'_{\text{identity}}(G; x, y, X, Y) = \\ \| G(x, Y, X) - x \|_1 + \\ \| G(y, X, Y) - y \|_1, \end{aligned} \tag{8}$$

여기서 $G(x, X, Y)$는 화자 X의 음성 데이터 x를 화자 Y의 음성 데이터로 변환하는 생성자로서 화자 식별 벡터 X와 Y를 조건으로 입력받고, $D(x, X)$는 음성 데이터 x가 화자 X의 음성인지 판별하는 구분자로서 화자 식별 벡터 X를 조건으로 입력받는다. 즉, CC-GAN에서 G와 D는 CycleGAN에서 G_X와 D_X 그리고 G_Y와 D_Y의 역할을 모두 수행할 수 있다. 따라서, 화자 식별 벡터 쌍의 개수를 늘리면 여러 개의 CycleGAN 효과를 내기 때문에 다수 화자 음성 변환에 활용할 수 있다.

2.2 CycleVAE

VAE는 인코더와 디코더로 구성된 생성 모델로서, 인코더를 통해서 입력 데이터를 잘 표현하는 잠재 변수를 찾아내고, 디코더를 통해서 잠재 변수를 원래 입력 데이터로 복원할 수 있다. VAE의 학습에는 Evidence Lower Bound(ELBO)가 활용되는데, 음성 변환에서는 화자 정보를 디코더에 조건으로 주는 다음과 같은 목적 함수를 학습에 사용한다.

$$\begin{aligned} L_{\text{VAE}}(\phi, \theta; x, X) = \\ \mathbb{D}_{\text{KL}}(q_\phi(z|x) \| p(z)) - \\ \mathbb{E}_{z \sim q_\phi(z|x)}[\log p_\theta(x|z, X)], \end{aligned} \tag{9}$$

여기서 x는 입력 음성 데이터, X는 화자 식별 벡터, $\mathbb{D}_{\text{KL}}$은 Kullback-Leibler divergence, $\mathbb{E}$는 기댓값, q_ϕ는 ϕ를 파라미터로 갖는 인코더, p_θ는 θ를 파라미터로 갖는 디코더, $p(z)$는 z의 사전 확률 분포, z는 잠재 변수를 의미한다. 학습이 완료된 후 음성 변환 시에는 X 대신에 목표 화자의 화자 식별 벡터를 디코더에 조건으로 줌으로써 입력 화자 음성이 목표 화자 음성으로 변환된다.

기존의 VAE는 입력 데이터를 잠재 변수로 표현한

후 이것을 다시 입력 데이터로 복원하는 것을 학습하기 때문에 원본 화자 음성을 목표 화자 음성으로 변환하는 과정을 직접적으로 학습하는 것은 아니다. 따라서, 기존의 VAE 기반 음성 변환은 직접적인 변환 과정을 학습하는 GAN 기반 음성 변환에 비해서 일반적으로 낮은 성능을 보인다. 이러한 단점을 보완하기 위해서 CycleVAE는 직접적인 변환 과정과 관련된 cycle consistency loss를 ELBO를 활용하여 추가한다. 화자 X의 음성 데이터 x와 화자 Y의 음성 데이터 y가 주어졌을 때 CycleVAE의 목적 함수는 다음과 같다.

$$L_{\mathrm{CycleVAE}}(\phi,\theta;x,y,X,Y) = L_{\mathrm{VAE}}(\phi,\theta;x,X) + L_{\mathrm{VAE}}(\phi,\theta;y,Y) + \lambda_3 L_{\mathrm{cydeELBO}}(\phi,\theta;x,X,Y) + \lambda_3 L_{\mathrm{cydeELBO}}(\phi,\theta;y,Y,X), \tag{10}$$

$$L_{\mathrm{cydeELBO}}(\phi,\theta;x,X,Y) = \mathbb{D}_{\mathrm{KL}}(q_\phi(z|x'_{X\to Y}) \parallel p(z)) - \mathbb{E}_{z\sim q_\phi(z|x'_{X\to Y})}[\log p_\theta(x|z,X)], \tag{11}$$

여기서 $x'_{X\to Y}$는 화자 X의 음성 데이터 x가 CycleVAE를 통해서 화자 Y의 목소리로 변환된 음성 데이터, λ_3은 $L_{\mathrm{cycleELBO}}$의 상대적인 가중치를 의미한다.

III. 실 험

CC-GAN과 CycleVAE의 화자 수에 따른 성능을 비교 분석하기 위하여 한국어 음성 데이터를 사용하여 다양한 실험을 하였다.

3.1 실험 환경

실험에는 국립국어원에서 배포한 서울말 낭독체 발화 말뭉치를 사용하였다. 임의로 4, 10, 100명을 추출하여 각각 다수 화자 음성 변환 성능을 측정하였다. 4명의 화자(fv01, fy03, mv01, mv02) 데이터는 10명의 데이터에 포함되게 하고, 10명의 데이터는 100명의 데이터에 포함되게 하였으며, 남녀의 비율은 균등하게 구성하였고, 화자 당 200 문장의 학습 데이터와 15 문장의 테스트 데이터를 사용하였다. 학습 화자 수의 변화에 따른 성능을 공정하게 비교하기 위하여 모두 공통인 4명의 테스트 데이터를 기준으로 음성 변환 성능을 측정하였다.

음성 데이터 파일에서 36 차원의 Mel-Frequency Cepstral Coefficient(MFCC), aperiodicity, logarithmic fundamental frequency를 매 5 ms마다 추출하였다. 학습에는 128 프레임 단위의 MFCC 벡터를 입력으로 사용하였다. 음성 변환 시에는 II 장에서 언급한 방법에 의해서 변환된 MFCC, logarithm Gaussian normalized transformation으로 변환된 logarithmic fundamental frequency, 그리고 원본 화자의 aperiodicity를 이용해서 음성 파형을 생성하였다.[21]

실험에 사용한 CC-GAN과 CycleVAE의 모델 구조와 하이퍼파라미터는 기존 연구를 참고하였다.[11,19] CC-GAN과 CycleVAE의 파라미터 개수는 각각 161,952와 46,562이었다. CC-GAN 학습 시, 10 화자의 경우 생성자에는 0.0001, 구분자에는 0.0005의 학습률을 적용하였고, 100 화자의 경우 생성자에는 0.00002, 구분자에는 0.00001의 학습률을 적용하였다. L'_{GAN}, L'_{cycle}, L'_{identity}의 상대적인 가중치 비율은 3:10:5로 설정하였고, L'_{identity}는 전체 에포크 횟수 중에서 초반 1/4에만 사용하였다. 4 화자 모델의 경우 1,800 에포크까지 학습을 진행하였고, 10 및 100 화자 모델의 경우 시간을 기준으로 하였다(3.3절에서 설명). CycleVAE의 경우 0.001의 학습률을 적용하였고, λ_3을 0.5로 설정하였으며, 4 화자 모델의 경우 2,000 에포크까지 학습하였다. 모든 실험에서 Adam 옵티마이저로[22] 학습을 진행하였고, 전체 에포크 수의 1/2이 지난 시점에서 선형적으로 학습률을 감소하였다.

음성 변환의 객관적인 성능 측정을 위해서 Mel-Cepstral Distortion(MCD)과[23] Modulation Spectra Distance(MSD)[24] 등의 정량적 수치를 사용하였고, Modulation Spectrum(MS) 그래프를 분석하였다. MCD와 MSD는 작을수록 목표 화자 음성과 유사하다는 것을 의미하고, MS 그래프는 높게 그려질수록 변환된 음성의 지나친 평활화가 적게 일어난다는 것을 의미한다.

3.2 소규모 화자 음성 변환 실험

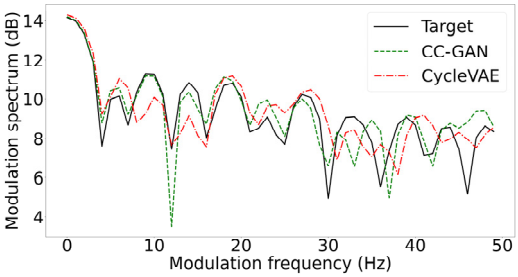
4 화자 음성 변환 실험의 경우 NVIDIA Quadro RTX 6000에서 CC-GAN 모델 학습에는 약 130 h, CycleVAE 모델 학습에는 약 4.6 h가 소요되었다. Table 1과 Table 2는 CC-GAN과 CycyleVAE로 각각 변환된 음성의 MCD 평균값과 신뢰구간 그리고 MSD 평균값과 신뢰구간을 각각 보여준다. CC-GAN이 CycleVAE보다 우수한 MCD 및 MSD 값을 갖는다는 것을 알 수 있다.

Table 1. The mean MCD and 95 % confidence interval of the voice conversions among 4 speakers. 'F' and 'M' represent female and male, respectively.

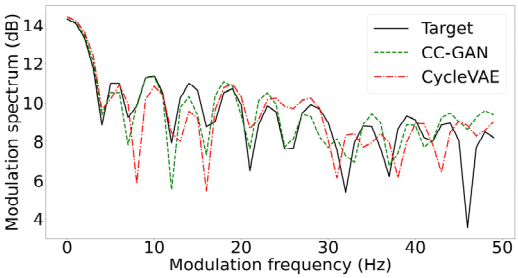
	CC-GAN	CycleVAE
F to F	6.522 ± 0.225	6.614 ± 0.207
M to F	6.430 ± 0.151	6.939 ± 0.162
F to M	6.161 ± 0.126	6.358 ± 0.130
M to M	5.816 ± 0.161	6.098 ± 0.128
Average	6.253 ± 0.086	6.550 ± 0.091

Table 2. The mean MSD and 95 % confidence interval of the voice conversions among 4 speakers.

	CC-GAN	CycleVAE
F to F	1.770 ± 0.009	2.028 ± 0.034
M to F	1.765 ± 0.010	2.093 ± 0.026
F to M	1.751 ± 0.014	2.003 ± 0.026
M to M	1.738 ± 0.029	2.054 ± 0.054
Average	1.757 ± 0.008	2.045 ± 0.017



(a) Modulation spectrum of 10th MFCC



(b) Modulation spectrum of 20th MFCC

Fig. 1. (Color available online) The modulation spectra of a sample utterance (t09_s08) by a target speaker (mv01) and two converted speech utterances from a source speaker (fv01) to the target speaker by CC-GAN and CycleVAE, respectively.

Fig. 1은 목표 화자 음성, CC-GAN에 의해서 변환된 음성, CycleVAE에 의해서 변환된 음성의 MS 그래프를 보여준다. CC-GAN 결과가 CycleVAE의 결과보다 목표 화자 음성과 조금 더 유사한 양상을 보인다는 것을 알 수 있다.

주관적인 듣기 평가에는 각 변환 방향 당 5 문장을 사용하였다. 듣기 평가에 참여한 평가자는 9 명이었다. 음질 듣기 평가의 경우 Mean Opinion Score(MOS)를 측정하였다. CC-GAN에 의해서 변환된 음성 20 문장, CycleVAE에 의해서 변환된 음성 20 문장, 목표 화자 음성 20 문장을 무작위 순서로 평가자에게 들려주고 음성 품질을 1(매우 나쁨), 2(나쁨), 3(보통), 4(좋음), 5(매우 좋음) 중에 하나로 평가하게 하였다. Table 3은 음질 듣기 평가의 MOS 평균과 신뢰구간을 보여준다. CC-GAN이 CycleVAE보다 높은 MOS 수치를 갖는다는 것을 알 수 있다.

유사도 듣기 평가에서는 목표 화자의 음성을 먼저 들려주고 CC-GAN과 CycleVAE에 의해서 변환된 음

Table 3. The mean MOS and 95 % confidence interval of the sound quality test.

	CC-GAN	CycleVAE	Target speaker
F to F	2.911 ± 0.853	2.089 ± 0.688	4.222 ± 0.351
M to F	3.244 ± 0.805	1.911 ± 0.615	
F to M	3.489 ± 0.769	1.867 ± 0.570	4.589 ± 0.135
M to M	3.733 ± 0.657	2.178 ± 0.711	
Average	3.344 ± 0.223	2.011 ± 0.131	4.406 ± 0.191

Table 4. Similarity test result with 95% confidence interval (%).

	CC-GAN	Fair	CycleVAE
F to F	40.0 ± 18.8	44.4 ± 28.5	15.6 ± 12.8
M to F	57.8 ± 19.5	28.9 ± 19.0	13.3 ± 13.3
F to M	31.1 ± 21.9	60.0 ± 23.1	8.9 ± 11.2
M to M	60.0 ± 18.8	28.9 ± 21.9	11.1 ± 8.1
Average	47.2 ± 10.8	40.6 ± 10.0	12.2 ± 5.6

성을 무작위 순서로 들려준 후 둘 중에서 어느 것이 목표 화자 음성과 더 비슷한지 선택하게 하였다. 어느 하나를 선택하기 어려운 경우 'Fair'를 선택할 수도 있었다. Table 4는 유사도 듣기 평가의 결과를 보여준다. CC-GAN으로 변환한 음성이 CycleVAE로 변환한 음성보다 목표 화자의 음성과 더 비슷하다고 선택한 것을 알 수 있다.

3.3 대규모 화자 음성 변환 실험

소규모 화자의 음성 변환에는 CC-GAN이 우수한 성능을 보여주지만, 화자 수가 증가하면 모든 화자 쌍 데이터를 학습해야 하기 때문에 학습 시간이 지수적으로 증가할 가능성 있다. 본 논문에서는 화자 수 증가와 학습 시간 및 음성 변환 성능을 분석하기 위해서 화자 수를 4명에서 10, 100명으로 증가시켜 가면서 학습 시간에 따른 음성 변환 성능을 측정하였다. Tables 5와 6은 각각 10, 100 화자 음성 변환에서 MCD의 평균값과 신뢰구간을 보여준다. CC-GAN의 경우 한 에포크 당 학습 시간은 10 화자 모델에서 약 0.7 h, 100 화자 모델에서 약 127.5 h가 소요되었다. CycleVAE의 경우 한 에포크 당 학습 시간은 10 화자 모델에서 약 54 s, 100 화자 모델에서 약 1.6 h가 소요되었다. 제한된 학습 시간에서 10 화자의 경우 CC-GAN이 우수한 성능을 보였지만, 100 화자의 경우 CycleVAE가 우수한 성능을 보였다.

Table 5. The mean MCD and 95 % confidence interval of the voice conversions among 10 speakers.

Training time (h)	CC-GAN	CycleVAE
100	6.932 ± 0.105	7.013 ± 0.082
200	6.513 ± 0.078	**7.003** ± 0.085
300	6.485 ± 0.088	7.025 ± 0.086
400	6.449 ± 0.086	7.035 ± 0.087
500	**6.431** ± 0.085	7.013 ± 0.083

Table 6. The mean MCD and 95 % confidence interval of the voice conversions among 100 speakers.

Training time (h)	CC-GAN	CycleVAE
100	**8.066** ± 0.128	7.093 ± 0.091
200	8.264 ± 0.106	7.048 ± 0.088
300	8.177 ± 0.105	**7.040** ± 0.086
400	8.390 ± 0.108	7.040 ± 0.086
500	8.455 ± 0.097	7.046 ± 0.088

IV. 결 론

충분한 저장 공간과 병렬 학습이 지원된다면 모든 화자 쌍에 대하여 각각 음성 변환 시스템을 구축할 수도 있겠지만, 그렇지 않은 컴퓨팅 환경에서는 단일 시스템으로 다수 화자 음성 변환을 하여야 한다. 본 연구에서는 비병렬 학습 데이터를 활용하는 다수 화자 음성 변환 방법 중에서, 우수한 음성 품질을 보여주는 CC-GAN과 빠른 학습 시간을 보여주는 CycleVAE를 실험적으로 비교 분석하였다. CC-GAN은 CycleGAN에 화자 정보를 조건으로 제공하여 다수 화자의 음성 변환을 가능하게 하지만, 많은 학습 시간과 까다로운 수렴 조건을 가지는 GAN의 특성상 대규모 화자로의 확장에 어려움이 있었다. 반면에 CycleVAE는 약 28배에서 80배의 빠른 수렴 속도로 인하여 제한된 시간에서 대규모 화자로의 확장이 용이하였다. 즉, 4 화자에서 100 화자로 확장함에 따라서 MCD 기준으로 CC-GAN의 경우 29.0 %의 음성 왜곡 증가율을 보인 반면, CycleVAE는 7.5 %의 음성 왜곡 증가율을 보였다. 결과적으로 소규모 화자의 경우 CC-GAN이 CycleVAE에 비하여 4.5 % 적은 변환 음성의 왜곡을 보인 반면, 대규모 화자의 경우 CycleVAE가 CC-GAN에 비하여 12.7 % 적은 변환 음성의 왜곡을 보인다는 것을 실험적으로 밝혔다.

감사의 글

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2017R1E1A1A01078157).

References

1. B. Ko, K. Lee, I.-C. Yoo, and D. Yook, "Korean voice conversion experiments using CC-GAN and VAW-GAN" (in Korean), Proc, Speech Communication and Signal Processing, **36**, 39 (2019).
2. B. Jang, H. Seo, I.-C. Yoo, and D. Yook, "CycleVAE based many-to-many voice conversion experiments

using Korean speech corpus" (in Korean), J. Acoust. Soc. Suppl.2(s) **40**, 79 (2021).
3. I.-C. Yoo, K. Lee, S.-G. Leem, H. Oh, B. Ko, and D. Yook, "Speaker anonymization for personal information protection using voice conversion techniques," IEEE Access, **8**, 198637-198645 (2020).
4. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Proc. NIPS, 2672-2680 (2014).
5. D. Kingma and M. Welling, "Auto-encoding variational Bayes," arXiv:1312.6114 (2013).
6. J. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-to image translation using cycle-consistent adversarial networks," Proc. IEEE Int. Conf. Computer Vision, 2242-2251 (2017).
7. T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," Proc. EUSIPCO, 2114-2118 (2018).
8. T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion," Proc. IEEE ICASSP, 6820-6824 (2019).
9. T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC3: Examining and improving CycleGAN-VCs for Mel-spectrogram conversion," Proc. Interspeech, 2017-2021 (2020).
10. D. Yook, I.-C. Yoo, and S. Yoo, "Voice conversion using conditional CycleGAN," Proc. Int. Conf. CSCI, 1460-1461 (2018).
11. S. Lee, B. Ko, K. Lee, I.-C. Yoo, and D. Yook, "Many-to-many voice conversion using conditional cycle-consistent adversarial networks," Proc. IEEE ICASSP, 6279-6283 (2020).
12. H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," Proc. IEEE Workshop on SLT, 266-273 (2018).
13. T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "StarGAN-VC2: Rethinking conditional methods for StarGAN-based voice conversion," Proc. Interspeech, 679-683 (2019).
14. C. Hsu, H. Hwang, Y. Wu, Y. Tsao, and H. Wang, "Voice conversion from non-parallel corpora using variational autoencoder," Proc. APSIPA, 1-6 (2016).
15. A. Oord and O. Vinyals, "Neural discrete representation learning," Proc. NIPS, 6309-6318 (2017).
16. C. Hsu, H. Hwang, Y. Wu, Y. Tsao, and H. Wang, "Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks," Proc. Interspeech, 3364-3368 (2017).
17. H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder," IEEE/ACM Trans. on Audio, Speech, and Lang. Process. **27**, 1432-1443 (2019).
18. P. Tobing, Y. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Non-parallel voice conversion with cyclic variational autoencoder," Proc. Interspeech, 674-678 (2019).
19. D. Yook, S.-G. Leem, K. Lee, and I.-C. Yoo, "Many-to-Many voice conversion using cycle-consistent variational autoencoder with multiple decoders," Proc. Odyssey: The Speaker Language Recognition Workshop, 215-221 (2020).
20. B. Ko, *Many-to-many voice conversion using cycle-consistency for Korean speech* (in Korean), (Master Thesis, Korea University, 2020).
21. M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," IEICE Trans. on Information and Systems, **99**, 1877-1884 (2016).
22. D. Kingma and J. Ba, "Adam: A method for stochastic optimization," Proc. ICLR, 1-13 (2015).
23. T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," IEEE Trans. on Audio, Speech, and Lang. Process. **15**, 2222-2235 (2007).
24. S. Takamichi, T. Toda, A. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," IEEE/ACM Trans. on Audio, Speech, and Lang. Process. **24**, 755-767 (2016).

저자 약력

▸ 육 동 석 (Dongsuk Yook)

1990년 8월 : 고려대학교 컴퓨터학과 학사
1993년 2월 : 고려대학교 컴퓨터학과 석사
1999년 8월 : Rutgers University, Department of Computer Science, Ph.D.
2001년 3월 ~ 현재 : 고려대학교 컴퓨터학과 교수

▸ 서 형 진 (HyungJin Seo)

2017년 3월 ~ 현재 : 고려대학교 컴퓨터학과 학사과정

▸ 고 봉 구 (Bonggu Ko)

2013년 2월 : 중앙대학교 문헌정보학과 학사
2020년 2월 : 고려대학교 컴퓨터학과 석사
2021년 8월 ~ 현재 : (주)위드캣 연구원(고려대학교 재학 기간에 이 논문에 기여하였음)

▸ 유 인 철 (In-Chul Yoo)

2006년 2월 : 고려대학교 컴퓨터학과 학사
2008년 2월 : 고려대학교 컴퓨터학과 석사
2015년 8월 : 고려대학교 컴퓨터학과 박사
2018년 2월 ~ 현재 : 고려대학교 컴퓨터학과 연구교수