CycleVAE를 이용한 다수 화자 한국어 음성 변환 실험

장봉원, 서형진, 유인철, 육동석

고려대학교 컴퓨터학과 인공지능 연구실

INTRODUCTION

원본 화자(source speaker)가 전달하려는 내용, 감정, 억양 등의 언어 정보를 유지하면서 원본 화자의 목소리를 목표 화자(target speaker)의 목소리로 변환하는 것을 음성 변환(voice conversion)이라고 한다. 이러한 음성 변환 기술은 발음 보조, 익명화를 통한 개인 정보 보호 등 다양한 영역에서 활용될 수 있다. 최근에는 DNN(deep neural network) 기반 음성 변환 연구가 활발히 진행되고 있다. 대표적으로 GAN(generative adversarial network)[1]과 VAE(variational autoencoder)[2]를 이용한 방법이 있다. 이러한 방법들은 음성 변환 모델을 학습시킴에 있어서 원본 화자와 목표 화자가 동일한 내용을 발화한 병렬 학습 데이터를 요구하지 않으므로 많은 학습 데이터를 활용할 수 있다는 장점이 있다. 특히, VAE 기반 음성 변환 모델 가운데 목표 화자 정보를 디코더의 조건으로 사용하는 conditional VAE가 다수 화자 간 음성 변환 학습에서 효율성을 보여주고 있다. 그러나 conditional VAE 기반의 음성 변환은 변환된 음성의 음질이 낮다는 단점이 있다. 본 연구에서는 conditional VAE의 단점을 보완하기 위해 명시적 변환 경로 학습 과정을 추가하여 변환된 음성의 품질을 개선한 CycleVAE(cycle-consistent variational autoencoder)를 한국어 음성 변환 실험에 적용한다. 기존의 연구에서는 적은 수의 화자들이 실험에 사용되었으나, 본 연구에서는 100명의 한국어 화자 데이터를 사용하여 CycleVAE의 음성 변환 성능을 실험한다.

RELATED WORKS

VAE는 인코더(encoder)와 디코더(decoder)로 구성된 생성 모델이다. VAE의 인코더는 입력 데이터를 잠재 변수(latent variable)로 표현하며, 디코더는 잠재 변수를 원래 입력 데이터로 복원한다. 음성 변환의 경우엔 화자 정보를 디코더에 조건으로 주는 것으로 아래와 같은 목적 함수를 사용하는 conditional VAE가 사용된다.

$$\mathcal{L}_{\mathrm{VAE}}(\Phi,\theta;x,A)=\mathbb{D}_{\mathrm{KL}}\big(q_\Phi(z|x)||p(z)\big)-\mathbb{E}_{z\sim q_\Phi(z|x)}\big[\log\big(p_\theta(x|z,A)\big)\big]$$

$\mathbb{D}_{\mathrm{KL}}$은 Kullback-Leibler divergence, q_Φ는 Φ를 파라미터로 갖는 인코더, p_θ는 θ를 파라미터로 갖는 디코더, $p(z)$는 z의 사전 확률 분포(prior probability distribution), x는 입력 데이터, z는 잠재 변수를 의미한다. 여기서 A는 화자 식별 벡터로서 음성 데이터 x를 발화한 화자를 가리킨다. 학습이 완료된 후 음성 변환 시에는 A 대신에 목표 화자의 화자 식별 벡터를 디코더에 조건으로 주는 것으로 입력 화자 음성이 목표 화자 음성으로 변환된다. 기존의 VAE는 입력 데이터를 잠재 변수로 표현한 후 다시 입력 데이터로 복원하는 방식으로 학습하므로, 원본 화자 음성을 목표 화자 음성으로 변환하는 과정을 명시적으로 학습하지 않는다. 따라서, 기존의 VAE 기반 음성 변환은 일반적으로 낮은 성능을 보인다.

PROPOSED METHOD

CycleVAE[3]는 변환 과정을 명시적으로 학습하기 위해 cycle consistency loss를 목적 함수에 추가한다. CycleVAE는 하나의 디코더에 목표 화자 식별 벡터를 조건으로 주거나 목표 화자마다 디코더를 할당하는 형태로 구현될 수 있는데, 본 실험에서는 목표 화자 식별 벡터를 디코더에 조건으로 주는 형태로 CycleVAE를 구현하였다. 화자 A의 음성 데이터 x와 화자 B의 음성 데이터 y가 주어질 때, CycleVAE의 목적 함수는 아래와 같다.

Equation 1: CycleVAE loss function

$$\mathcal{L}_{\mathrm{CycleVAE}}(\Phi,\theta;x,y,A,B)=\mathcal{L}_{\mathrm{VAE}}(\Phi,\theta;x,A)+\mathcal{L}_{\mathrm{VAE}}(\Phi,\theta;y,B)+\lambda\cdot\mathcal{L}_{\mathrm{cycle}}(\Phi,\theta;x,A,B)+\lambda\cdot\mathcal{L}_{\mathrm{cycle}}(\Phi,\theta;y,B,A)$$

Equation 2: cycle consistency loss function

$$\mathcal{L}_{\mathrm{cycle}}(\Phi,\theta;x,A,B)=\mathbb{D}_{\mathrm{KL}}(q_\Phi(z|x'_{A\to B})\,\|\,p(z))-\mathbb{E}_{z\sim q_\Phi(z|x'_{A\to B})}[\log p_\theta(x|z,A)]$$

Equation 1에서 λ는 cycle consistency loss의 가중치를 나타낸다. CycleVAE loss function은 화자 A의 음성 데이터 x가 주어질 때, 인코더와 디코더를 통해 똑같은 화자 A의 음성 데이터 x로 복원하면서 계산하는 VAE loss function과 화자만 B로 바꿨다가 다시 원래 화자 A로 복원하면서 계산하는 cycle consistency loss function으로 이루어져 있다. Equation 2에서 $x'_{A\to B}$은 화자 A가 발화한 음성 데이터 x를 화자 B의 목소리로 음성 변환한 것을 가리킨다.

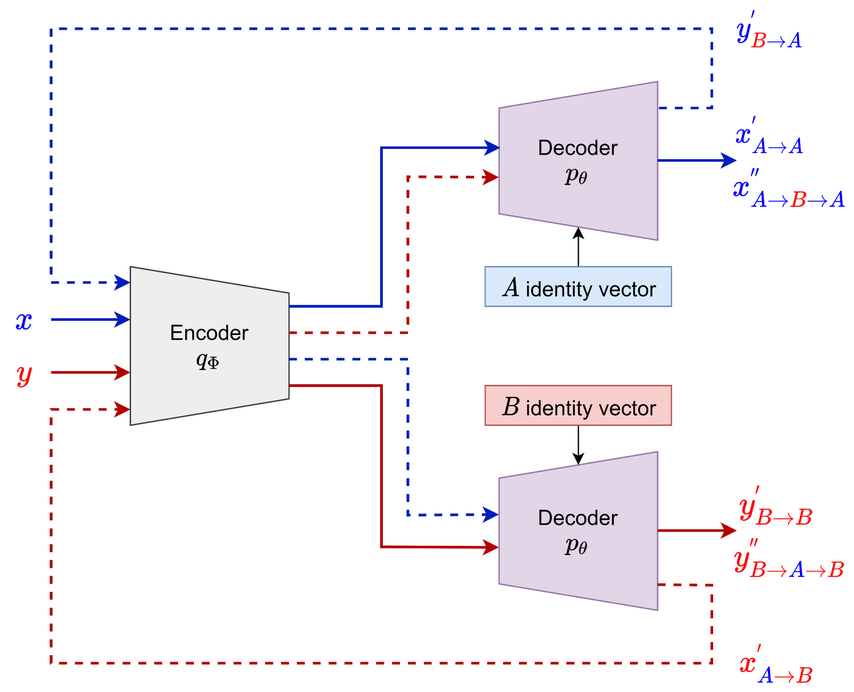

Figure 1: The architecture of CycleVAE for two speakers. In this figure, a decoder is drawn for each target speaker for ease of explanation. In the actual experiments, only one decoder is used, and the target speaker identity vector is used as a condition for voice conversion.

EXPERIMENTS

CycleVAE의 성능을 확인하기 위해 한국어 음성 변환 실험에 적용했다. 실험에는 서울말 낭독체 말뭉치를 사용했다. 10, 100명의 화자에 대하여 다수 화자 음성 변환 성능을 측정했다. 남녀의 비율은 균등하며, 화자 당 200 문장의 학습 데이터와 15 문장의 테스트 데이터를 사용했다. 음성 변환의 객관적인 성능 확인을 위해서 정량적 수치인 MCD(Mel-cepstral distortion) 값을 사용했다.

10 Speakers	
Training time (hour)	MCD
0	8.40±0.033
2	6.75±0.028
4	6.72±0.029
6	6.70±0.029
8	6.70±0.028
10	6.68±0.029
20	6.67±0.029
100	6.65±0.029

100 Speakers	
Training time (hour)	MCD
0	8.50±0.029
2	6.81±0.028
4	6.69±0.026
6	6.95±0.027
8	6.72±0.029
10	6.61±0.026
20	6.86±0.037
100	6.62±0.028

Table 1: Mean MCD (Mel-cepstral distortion) of the voice conversion among 10 speakers and 100 speakers, respectively.

Table 1은 10, 100 화자 음성 변환에서 학습 시간에 따른 MCD의 평균과 표준편차를 보여준다. 10, 100화자 변환 실험에서 모두 제한된 학습 시간 내에서 우수한 성능을 보였으며, 특히 100화자로 화자 수가 늘었어도 CycleVAE를 통해 변환된 음성의 품질이 유지되는 것을 확인할 수 있었다.

CONCLUSIONS

비병렬 학습 데이터를 사용해서 CycleVAE를 통한 다수 화자 음성 변환 성능을 확인했다. 기존의 conditional VAE와 달리, 명시적인 변환 과정을 학습할 수 있도록 개선한 CycleVAE는 10, 100명의 화자에 대한 다수 음성 변환 실험에서 우수한 성능을 보였다.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” Neural Information Processing Systems, pp. 2672-2680 (2014)

[2] D. Kingma, and M. Welling, “Auto-Encoding Variational Bayes,” arXiv:1312.6114 (2014)

[3] D. Yook, S.-G. Leem, K. Lee, and I.-C. Yoo, “Many-to-Many Voice Conversion using Cycle-Consistent Variational Autoencoder with Multiple Decoders,” Odyssey 2020: The Speaker and Language Recognition Workshop, pp. 215-221 (2020)