마스크 착용에 의해 왜곡된 음성의 품질 향상을 위한 CycleGAN 기술

*임유진 유정찬 서은미 박호종
광운대학교
*dbwls553@naver.com

CycleGAN for Enhancement of Degraded Speech by Face Mask

*Lim, Yujin Yu, Jeongchan Seo, Eunmi Park, Hochong
Kwangwoon University

요약

마스크 착용은 대화나 통화 등의 의사소통에 불편함을 초래하고 음성의 품질과 명료도를 떨어트린다. 이를 해결하기 위해 음성 향상 기술이 필요하며, 머신러닝 기반의 다양한 음성 향상 방법이 개발되었다. 지도 학습을 위해 마스크 착용 유무에 따라 일대일로 대응된 음성 데이터를 확보하는 것은 매우 어렵고, 따라서 일대일로 대응된 데이터가 필수적이지 않은 비지도 학습이 요구된다. 본 논문에서는 비지도 학습방식을 사용하면서 콘텍스트를 유지하며 특징을 변경할 수 있는 CycleGAN을 이용하여 마스크 착용에 의한 음성 왜곡을 복원 시키는 기술을 제안한다. 스펙트로그램 기반으로 마스크 착용에 의해 왜곡된 음성을 마스크 미착용 음성으로 변환하여 음성의 품질을 향상시켰다. 청취평가를 진행한 결과 품질이 향상된 음원의 선호도가 더 높음을 확인하였으며 스펙트로그램을 통해 3 kHz 이상의 고대역 에너지가 증가하는 것을 확인하였다. 이를 통해 CycleGAN을 이용한 비지도 학습으로 마스크 착용에 의해 왜곡된 음성의 품질을 향상시킬 수 있음을 확인하였다.

1. 서론

최근 COVID-19로 인해 마스크 착용이 보편화되었다. 마스크는 음성의 품질을 저하시키고 안면근육의 움직임을 방해하거나 안면의 인식을 어렵게 하여 대화, 통화 등의 의사소통에 불편함을 초래한다. 특히 청각적인 측면에서 3 kHz 이상의 고대역 에너지를 감소시켜 파열음, 마찰음과 같은 무성음의 명료도를 떨어트려 인식을 어렵게 한다[1]. 또한, 음성인식과 같이 머신러닝을 이용한 음성 신호처리에서도 마스크 착용에 의한 음성의 품질저하는 모델의 성능 하락을 야기한다. 이러한 문제를 해결하기 위해 본 논문에서는 CycleGAN을 이용하여 마스크 착용에 의해 왜곡된 음성의 품질을 향상시키는 기술을 제안한다. 제안하는 기술은 COVID-19 상황에서뿐만 아니라 산업현장, 의료분야 등 직업 특성상 마스크의 착용이 필수한 곳에서도 이용될 수 있다.

일반적으로 음성 향상을 위한 머신러닝 모델은 일대일로 대응된 학습 데이터를 사용하여 지도 학습 방식으로 훈련을 진행한다. 하지만 마스크 착용 유무에 따라 일대일로 대응된 음성 데이터 녹음과 같이 통제 변인을 동일하게 유지할 수 없는 경우 일대일로 대응된 데이터를 확보하는 것이 매우 어렵다. 따라서 본 논문에서는 일대일로 대응된 데이터 없이도 콘텍스트(context)는 유지하며 특징을 변경할 수 있는 CycleGAN을 이용한 음성 향상 기술을 제안한다[2].

본 논문에서 제안한 CycleGAN 기반 음성 향상 기술의 성능을 A/B 선호도 검사(preference test)로 평가하였으며, 제안한 방법으로 생성한 음성의 선호도가 마스크 착용에 의해 왜곡된 음성보다 높은 것을 확인하였다. 또한 CycleGAN 기반 음성 향상에 의하여 3 kHz 이상의 고대역 에너지가 증가하는 것을 스펙트로그램으로 확인하였다. 이를 통해 CycleGAN을 이용하여 비지도 학습으로 마스크에 의해 왜곡된 음성을 향상시킬 수 있음을 확인하였다.

2. 제안하는 방법

본 논문에서는 비지도 학습 방식의 CycleGAN을 이용하여 마스크 착용에 의해 왜곡된 음성의 품질을 향상시키는 기술을 제안한다. 스펙트로그램 기반으로 마스크 착용에 의해 왜곡된 음성을 마스크 미착용 음성으로 변환하여 음성의 품질을 향상시키며, 위상에는 별도의 변형을 가하지 않는다. 모델에 입력되는 스펙트로그램은 16 kHz로 샘플링 된 음성 신호에 프레임 길이 1024 샘플, 50% 중첩, hanning 윈도우를 적용하여 short time fourier transform (STFT)하고 절대 값 및 로그를 취하여 구한다.

제안하는 방법은 그림 1을 통해 간략히 나타낼 수 있다. 마스크 착용 음성의 스펙트로그램 X를 생성기(generator) G에 입력하여 마스크 미착용 음성의 스펙트로그램 $\hat{Y}$을 얻는다. $\hat{Y}$은 판별기(discriminator) D를 통해 목표 신호에 맞도록 스펙트로그램 특징이 잘 변경되었는지 평가된다. 이를 통해 G와 D는 적대적 학습을 통해 특징을 보다 잘 변경할 수 있도록 업데이트 된다. 이 동작과 동시에, 마스크 착용 음성의 스펙트로그램으로 복원 하는 생성기 F를 통해 $\hat{X}$으로 역 변환되도록 유도하여 콘텍스트를 유지할 수 있도록 한다.

생성기와 판별기의 구조는 그림 2와 같다. 생성기와 판별기는 convolutional neural network (CNN)으로 구성되었으며, 생성기는 스펙트로그램의 차이만 구하도록 잔차(residual) 신경망 구조로 구성하였다. 생성기 G와 F는 동일한 구조를 가진다.

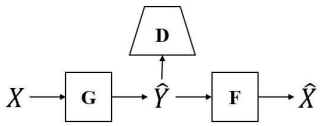

그림 1. 제안하는 모델의 구조
Fig. 1. Structure of proposed model

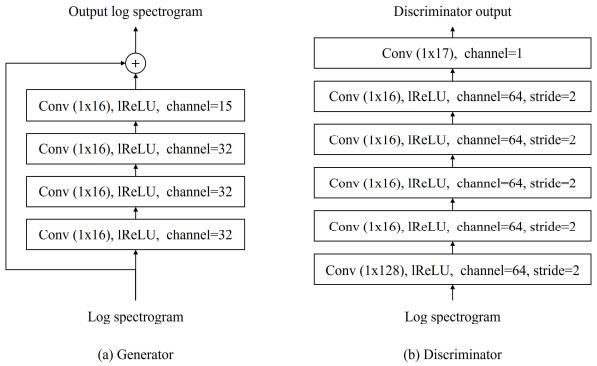

그림 2. 생성기와 판별기의 네트워크 구조
Fig. 2. Network architecture of generator and discriminator model

생성기와 판별기를 업데이트하기 위한 손실 함수는 각각 식 (1)과 식 (5)로 정의된다. L_G에서 L_{GAN-G}는 식 (2)로 정의되고, G의 결과인 $\hat{Y}$이 마스크 미착용 음성의 스펙트로그램으로 판단되도록 G를 적대적 방식으로 학습시킨다. L_{cyc}은 식 (3)으로 정의되고, $\hat{X}$이 X와 같아지도록 유도하여 $\hat{Y}$이 X의 콘텍스트를 유지하도록 생성기를 학습시킨다. 식 (4)에 정의된 L_{id}는 이미 특징이 변경 된 스펙트로그램이 입력됐을 때 입력과 같은 출력을 내보내도록 생성기를 학습시킨다. L_D는 적대적 손실 함수로만 정의된다. L_{GAN-G}와 L_D는 어떠한 적대적 손실 함수를 적용해도 무관하나 본 논문에서는 WGAN-gp의 손실 함수를 사용하였다[3]. L_D의 H는 WGAN-gp 손실 함수에서 gradient-penalty를 구하기 위한 벡터로 $G(X)$와 Y사이의 임의의 내분점을 의미한다. λ는 각각 손실 함수의 가중치를 조절하는 변수로 본 논문에서는 $\lambda_{cyc}=1$, $\lambda_{id}=1$, $\lambda_{gp}=10$으로 고정하였다.

$$L_G = L_{GAN-G} + \lambda_{cyc}L_{cyc} + \lambda_{id}L_{id} \quad (1)$$

$$L_{GAN-G} = -\mathbb{E}_{X\sim p_{data}(X)}[D(G(X))] \quad (2)$$

$$L_{cyc} = \mathbb{E}_{X\sim p_{data}(X)}[\|F(G(X))-X\|_1] + \mathbb{E}_{Y\sim p_{data}(Y)}[\|G(F(Y))-Y\|_1] \quad (3)$$

$$L_{id} = \mathbb{E}_{Y\sim p_{data}(Y)}[\|G(Y)-Y\|_1] + \mathbb{E}_{X\sim p_{data}(X)}[\|F(X)-X\|_1] \quad (4)$$

$$L_D = \mathbb{E}_{X\sim p_{data}(X)}[D(G(X))] - \mathbb{E}_{Y\sim p_{data}(Y)}[D(Y)] + \lambda_{gp}\mathbb{E}_{H\sim p_{data}(H)}[(\|\nabla_H D(H)\|_2 - 1)^2] \quad (5)$$

3. 성능 평가

훈련 및 성능 평가 데이터는 카이스트 오디오북 데이터 셋을 가공하고 재녹음하여 생성하였다. 한국어 혼성 10 명과 영어 혼성 2 명 등 총 12명을 선택하고, 각 화자 당 10분 씩 총 120분의 음성 데이터를 추출하여 16 kHz 샘플링 주파수로 재녹음 하였다. 성인 인간의 머리와 가슴의 음향 특성을 고려하여 음성을 실제와 같이 재현하도록 설계 된 head and torso simulator (HATS)를 이용하여 제한된 조건으로 완전 무향실에서 녹음하였다. 마스크는 KF94, KF80, KFAD의 일반형과 부리형 모델을 포함하여 총 6가지를 사용하였으며, 마스크를 착용하지 않은 경우를 포함하여 총 7번 녹음을 진행하였다. 6종류의 마스크 데이터 총 720분 중 648분을 훈련에 사용하였다. 훈련에 사용한 648분은 생성기 훈련에 324분, 판별기 훈련에 324분으로 일대일로 대응되지 않도록 분리하여 사용하였다.

A/B 선호도 검사를 위해 남자 3명(한국어 2명, 영어 1명), 여자 3명(한국어 2명, 영어 1명) 총 6명의 화자에 대하여 각각 15초의 음성 데이터를 선정하고, 6 종류의 마스크에 대해 평가를 진행하였다. 총 5명이 평가에 참여하여 품질 향상 전과 후의 음성에 대한 선호도를 비공개로 평가하였다. 표 1은 제안하는 방법으로 음성을 향상시키기 전과 후의 품질 선호도 선택 비율이고, 제안하는 방법을 사용하였을 때 더 높은 선호도를 가지는 것을 확인하였다. 그림 3은 제안하는 방법으로 음성을 향상시키기 전과 후의 스펙트로그램을 보여주며, 3 kHz 이상의 고대역 에너지가 증가하는 것을 확인하였다.

표 1. A/B 선호도 검사 결과
Table 1. Result of A/B preference test

Preference / Mask type	Before Enhancement	After Enhancement	Equivalent
KFAD	2.50 %	**90.42 %**	7.08 %
KF80	5.00 %	**89.58 %**	5.42 %
KF94	5.42 %	**85.83 %**	8.75 %

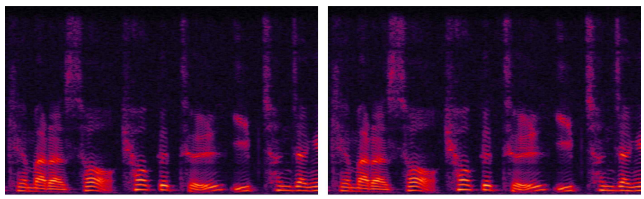

그림 3. 음성 향상 전과 후의 스펙트로그램
Fig. 3. Spectrogram before and after speech enhancement

4. 결론

본 논문은 CycleGAN을 이용하여 마스크 착용에 의해 왜곡된 음성의 품질을 향상시키는 기술을 제안하였다. 비지도 학습 방식으로 콘텍스트는 유지하며 왜곡된 음성의 품질 향상이 가능함을 성능평가를 통해 확인하였다.

감사의 글

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2021R1F1A1059233).

참고문헌

[1] M. Magee, *et al.*, “Effects of face masks on acoustic analysis and speech perception: Implications for peri-pandemic protocols,” *J. of Acoust. Soc. of America*, vol. 148, no. 6, pp. 3562-3568, 2020.

[2] J. Y. Zhu, *et al.*, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 2223-2232, 2017.

[3] I. Gulrajani, *et al.*, "Improved Training of Wasserstein GANs," *Advances in Neural Information Processing Systems (NIPS)*, pp. 5767-5777, 2017.