KIPS Trans. Softw. and Data Eng.
Vol.12, No.7 pp.303~314
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2023.12.7.303

Extending StarGAN-VC to Unseen Speakers Using RawNet3 Speaker Representation

Bogyung Park† · Somin Park†† · Hyunki Hong†††

ABSTRACT

Voice conversion, a technology that allows an individual's speech data to be regenerated with the acoustic properties(tone, cadence, gender) of another, has countless applications in education, communication, and entertainment. This paper proposes an approach based on the StarGAN-VC model that generates realistic-sounding speech without requiring parallel utterances. To overcome the constraints of the existing StarGAN-VC model that utilizes one-hot vectors of original and target speaker information, this paper extracts feature vectors of target speakers using a pre-trained version of Rawnet3. This results in a latent space where voice conversion can be performed without direct speaker-to-speaker mappings, enabling an any-to-any structure. In addition to the loss terms used in the original StarGAN-VC model, Wasserstein distance is used as a loss term to ensure that generated voice segments match the acoustic properties of the target voice. Two Time-Scale Update Rule (TTUR) is also used to facilitate stable training. Experimental results show that the proposed method outperforms previous methods, including the StarGAN-VC network on which it was based.

Keywords : Voice Conversion, Speaker Attribute, Generalization, StarGAN-VC, RawNet3

RawNet3 화자 표현을 활용한 임의의 화자 간 음성 변환을 위한 StarGAN의 확장

박 보 경† · 박 소 민†† · 홍 현 기†††

요 약

음성 변환(Voice Conversion)은 개인의 음성 데이터를 다른 사람의 음향적 특성(음조, 리듬, 성별 등)으로 재생성할 수 있는 기술로, 교육, 의사소통, 엔터테인먼트 등 다양한 분야에서 활용되고 있다. 본 논문은 StarGAN-VC 모델을 기반으로 한 접근 방식을 제안하여, 병렬 발화(Utterance) 없이도 현실적인 음성을 생성할 수 있다. 고정된 원본(source) 및 목표(target)화자 정보의 원핫 벡터(One-hot vector)를 이용하는 기존 StarGAN-VC 모델의 제약을 극복하기 위해, 본 논문에서는 사전 훈련된 Rawnet3를 사용하여 목표화자의 특징 벡터를 추출한다. 이를 통해 음성 변환은 직접적인 화자 간 매핑 없이 잠재 공간(latent space)에서 이루어져 many-to-many를 넘어서 any-to-any 구조가 가능하다. 기존 StarGAN-VC 모델에서 사용된 손실함수 외에도, Wasserstein-1 거리를 사용하여 생성된 음성 세그먼트가 목표 음성의 음향적 특성과 일치하도록 보장했다. 또한, 안정적인 훈련을 위해 Two Time-Scale Update Rule (TTUR)을 사용한다. 본 논문에서 제시한 평가 지표들을 적용한 실험 결과에 따르면, 제한된 목소리 변환만이 가능한 기존 StarGAN-VC 기법 대비, 본 논문의 제안 방법을 통해 다양한 발화자에 대한 성능이 개선된 음성 변환을 제공할 수 있음을 정량적으로 확인하였다.

키워드 : 음성 변환, 화자 특성, 일반화, StarGAN-VC, RawNet3

1. 서 론

인간은 표정, 행동과 같은 단순한 표현 외에도 음성 및 문자와 같은 방법으로 정교하게 자기 생각을 전달하며 상호작용을 한다. 인공지능 기술이 발전하면서 인간의 다양한 표현 방법을 분석하고 모방하는 연구가 발전되어 가고 있으며, 특히 음성에 관한 연구[1,2]들도 활발히 제안되고 있다.

텍스트에 관련된 인공지능 기술에는 그림 또는 영상을 글로 설명해 주는 기술(image-to-text, video-to-text)[3,4]과 자동 음성 인식(Automatic Speech Recognition, ASR)[5,6] 등이 있다. 이와 반대로 텍스트를 음성으로 변환해 주는 기술인 TTS(Text-to-speech)[7,8]도 있다. 그러나 기존 TTS 방법은 텍스트를 음성으로 변환하는 데에만 초점을 맞춰 정형화된 목소리로만 출력한다. 그러나 가상 인간(Virtual human)[9]이나 오디오북 또는 시각 장애인의 화면 읽기 프로그램[10] 등 다양한 애플리케이션을 개발하기 위해서는 TTS 연구뿐 아니라, 여러 사람의 목소리를 지원하는 연구도 중요하다. 따라서 본 논문은 기존의 음성 변환의 제약을 극복하기 위해, 다양한 화자의 음성을 표현할 수 있는 any-to-any 방법의 StarGAN-VC[22] 모델을 제안한다.

화자 음성 변환 기술[11-13]은 발화 내용은 유지하면서, 입력된 음성을 다른 사람의 목소리로 변환시키는 기술이다.

† 준 회 원 : 중앙대학교 AI학과 석사과정
†† 준 회 원 : 중앙대학교 컴퓨터공학과 석사과정
††† 준 회 원 : 중앙대학교 소프트웨어학부 교수
Manuscript Received : April 28, 2023
First Revision : June 9, 2023
Accepted : June 19, 2023
* Corresponding Author : Hyunki Hong(honghk@cau.ac.kr)

초기에는 Gaussian Mixture Models(GMMs)[14]을 이용했으며, 기계학습(Machine learning) 기술이 발전하면서 컨볼루션 신경망(Convolution Neural Network, CNN)[15] 또는 순환신경망(Recurrent Neural Network, RNN)[16]을 활용한 연구들이 제안되었다. 최근에는 적대적 생성 신경망(Generative Adversarial Neural Network, GAN)[17]을 활용하는 연구가 다양하게 진행되고 있다.

GAN은 실제 데이터의 분포를 이용해 이와 유사한 가짜 데이터를 생성해 내어, 어떠한 것이 실제 데이터이고 생성된 데이터인지 구분할 수 없게 적대적으로 학습해 나아가는 신경망 모델이다. 가장 기본적인 GAN의 구조는 가짜 데이터를 생성하는 생성자(Generator)와 데이터의 진위를 판단하여 구분해 주는 구별자(Discriminator)로 구성된다. GAN 모델은 우수한 성과를 보였지만, 모드 붕괴(Mode collapse), 학습 불안정성, 하이퍼 파라미터 민감도, 레이블이 없는 데이터 문제 등의 단점이 존재한다.

이러한 단점을 해결하기 위해 조건부 적대적 신경망(Conditional GAN, CGAN)[18]과 순환 적대적 신경망(CycleGAN)[19]과 같은 변형된 형태의 GAN이 제안되었다. 조건부 적대적 신경망은 조건부 정보를 추가하여 레이블을 활용해 이미지를 만든다. 순환 적대적 신경망은 도메인 간 일관성 손실함수(cycle consistency loss)를 사용하여 레이블이 없는 데이터에서도 도메인 간 이미지 변환을 수행한다.

또한, WGAN-GP[20]는 의미 있는 손실함수 측정, 그래디언트 페널티(Gradient penalty, GP) 도입하여 안정적인 학습을 가능하게 하여 더 나은 품질의 이미지를 생성한다.

TTUR[21]은 생성자와 판별자를 서로 다른 학습률(learning rate)과 업데이트 주기로 학습시키는 방법이다. 일반적으로 판별자는 생성자보다 빠른 학습률로 업데이트되며, 이에 따라 판별자가 생성자보다 더욱 정교하게 학습된다. TTUR을 사용하면 생성자와 판별자의 동적 균형이 유지되어 학습이 안정화되고, 모드 붕괴 문제를 완화할 수 있다.

본 논문에서 활용한 StarGAN-VC는 CycleGAN을 기반으로 만들어진 StarGAN[23]의 음성 변환 방법이다. 이는 다양한 음성 변환 작업을 단일 모델로 통합하기 위해 발전된 GAN 기반 방법으로, 음성 변환 과정에서 원본화자의 발화 내용은 보존하면서 목표화자의 음성 특성으로 변환한다. 이때, 복수 화자의 음성 변환을 동시에 학습(many-to-many)할 수 있는 효율적인 구조로 되어 있으며, 레이블이 있는 학습 데이터를 사용하여 각 화자의 음성 특성을 학습함으로써 안정적인 학습과 높은 품질의 음성 변환을 가능하도록 한다.

이처럼 화자 음성 변환에는 다양한 접근법이 존재한다. 먼저, 병렬(parallel) 접근법[24]은 원본 화자와 목표화자가 같은 발화 내용을 포함하는 정렬된 문장을 사용하지만, 데이터셋 수집과 생성할 수 있는 문장에 한계가 있다는 단점이 있다. 이러한 문제를 해결하기 위해 비 병렬(non-parallel) 접근법[25]이 제안되었다. 즉, 원본 화자의 발화 문장과 목표화자의 발화 문장이 서로 다른 다양한 문장을 사용하여 학습함으로써, 원본 화자의 내용은 유지하면서 목표화자의 목소리로 변환할 수 있다.

또 다른 접근 방법인 one-to-one 변환[26]은 한 화자의 음성을 특정 다른 화자의 음성으로 변환하는 것으로 간단한 구조로 되어 있지만, 확장성이 제한된다. 반면에 many-to-many 변환[27]은 다양한 화자 조합을 지원하며 더 효율적이고 확장성이 높지만, 복잡한 모델 구조와 다양한 변환을 동시에 학습해야 한다. StarGAN-VC는 many-to-many 변환을 지원하는 GAN 기반 방법으로, 하나의 모델을 이용하여 다양한 음성 변환 작업을 효율적으로 수행할 수 있다. 그러나 원핫 벡터를 사용하여 변환 대상 발화자를 표현함으로써 일반화 능력이 제한된다. 즉, 변환할 수 있는 화자의 수가 제한된다는 단점이 존재한다.

한편, RawNet3[28]는 화자 식별(speaker identification)과 화자 검증(speaker verification)을 위해 제안되었다. 원시 음성 파형(Raw waveform)을 입력으로 받아 특징을 추출하는데, 뛰어난 화자 인식(speaker recognition)의 성능을 보였다.

본 논문은 StarGAN-VC 모델을 활용함으로써, 비 병렬, many-to-many 접근 방법을 택하며, 하나의 모델을 이용하여 다양한 음성 변환 작업을 수행한다. 또한, 일반화 능력이 제한되는 StarGAN-VC의 단점을 보완하기 위해 RawNet3을 이용하여 목표화자의 특징을 추출하여 속성으로 입력한다. 이를 통해 변환할 수 있는 화자의 수를 제한하지 않는다. 즉, 학습 시 활용하지 않았던 화자(unseen speaker)를 포함한 다양한 발화자에 대해 any-to-any 음성 변환이 가능하다.

제안된 논문의 주요 기여 항목은 아래와 같다.

첫째, 본 논문은 기존 StarGAN-VC에 입력되는 목표화자 정보를 원핫 벡터 대신 RawNet3로부터 추출한 목표화자의 특징 벡터를 속성으로 입력한다. 이를 통해 발화자에 대한 일반화 능력을 높이고, 다양한 발화자에 대응할 수 있는 any-to-any 음성 변환을 수행할 수 있다.

둘째, WGAN-GP에서 안정적인 학습을 위해 소개된 Earth Mover's Distance(Wasserstein-1 Distance)를 사용하여 GAN의 적대적 손실함수를 개선하였으며, 그래디언트 페널티를 도입하여 립시츠(Lipschitz) 조건을 충족시켰다. 또한, TTUR을 사용하여 모드 붕괴 문제를 방지하였다. 이를 통해 학습이 안정되었으며, 더 높은 성능을 달성하였다.

2. 관련 연구

2.1 WORLD Vocoder

WORLD Vocoder[29]는 음성 신호를 세 가지 주요 구성 요소인 스펙트럼(spectral envelope, SP), 기본 주파수(fundamental frequency, F0), 그리고 비주기성(aperiodicity, AP)으로 분해하여 처리함으로써 음성의 품질과 처리 속도를 더욱 향상한다.

스펙트럼은 음성 신호의 주파수 영역에서 에너지 분포를 나타내며, 이를 분석하기 위해 푸리에 변환(Fourier Transform)[30]을 사용하여 시간 도메인의 음성 신호를 주파수 도메인으로 변환한다. 이러한 변환 과정을 통해 합성할 음성의

품질이 결정된다. 기본 주파수는 음성의 톤 높이를 결정하는 중요한 요소이며, 자기 상관 함수(autocorrelation function)와 같은 다양한 기법을 사용하여 음성의 고유한 특징을 정확하게 추정하고 조절할 수 있어, 이를 통해 자연스러운 음성 합성이 가능해진다. 비주기성은 음성 신호에서 불규칙한 변동을 나타내며, 세 발화 주기 알고리즘(Three-band Periodicity Algorithm)을 사용하여 음성의 세부적인 특징을 조절할 수 있다.

WORLD Vocoder는 음성 신호를 위의 세 구성 요소로 나누고, 이를 독립적으로 분석하고 처리함으로써 원본 음성과 매우 유사한 고품질의 음성을 빠르게 처리하여 합성할 수 있다.

2.2 WGAN-GP

WGAN[31]을 개선한 WGAN-GP[20]는 적대적 생성 신경망의 훈련 성능을 개선하기 위해 제안되었다. 기존 GAN에서 발생할 수 있는 학습 불안정과 모드 붕괴 등과 같은 문제를 완화하고 생성 모델의 성능을 향상한다.

WGAN은 립시츠 연속성(Lipschitz continuity)을 가정한 Earth Mover's Distance (Wasserstein-1 Distance)를 사용하여 실제 데이터 분포와 생성된 데이터 분포 간의 거리를 최소화한다. 기존 연구에는 립시츠 연속성 조건을 충족시키기 위해 네트워크 가중치 제한(weight clipping)을 해야 했으나, WGAN-GP에서 그래디언트 페널티 방법을 도입하여 실제 데이터와 생성된 데이터 사이의 임의의 점에 대해 그래디언트의 크기를 1에 가깝게 유지함으로써 립시츠 조건을 보장한다.

이 과정에서 Equation (1)과 같은 손실함수를 사용한다.

$$L = E_{\tilde{x}\sim\ \mathbb{P}_g}[D(\tilde{x})] - E_{\tilde{x}\sim\ \mathbb{P}_r}[D(x)] + \lambda E_{\hat{x}\sim\ \mathbb{P}_{\hat{x}}}[(\| \nabla_{\hat{x}} D(\hat{x}) \|_2 - 1)^2] \quad (1)$$

여기서 D는 판별기, G는 생성기를 의미하며, $\mathbb{P}_r$은 데이터 분포, $\mathbb{P}_g$은 $\tilde{x} = G(z)$, $(z \sim P(z))$에 의해 정의된 모델의 분포이다.

$E_{\hat{x}\sim\ \mathbb{P}_{\hat{x}}}[(\| \nabla_{\hat{x}} D(\hat{x}) \|_2 - 1)^2]$는 그래디언트 페널티 부분으로 λ는 그래디언트 페널티 가중치이고, $\hat{x}$은 실제 데이터와 생성된 데이터 사이의 임의의 점이다.

2.3 StarGAN-VC

StarGAN-VC[22]는 다중화자 음성 변환 작업을 효율적이고 유연하게 수행하기 위해 하나의 통합 모델을 사용한다. 여러 화자 간의 음성 변환을 처리할 수 있어 개별 화자마다 별도의 모델을 구축할 필요가 없으며 다양한 화자 조합에 대해 통일된 모델을 적용할 수 있다.

StarGAN-VC는 병렬 데이터셋이 필요하지 않아 생성할 수 있는 문장의 한계가 없는 특징을 가지고 있다. 기존의 음성 변환 기술에서는 원본 화자와 목표화자의 발화 문장이 짝지어진 데이터셋이 필요했지만, StarGAN-VC는 이러한 제약을 극복했다. 학습 과정에서 원본 화자의 발화 문장과 목표화자의 발화 문장이 서로 다른 다양한 문장을 사용하여 모델을 학습함으로써, 원본 화자의 내용은 유지하면서 목표화자의 목소리로 변환할 수 있다.

생성기의 upsample, 판별기의 downsample 과정에서 목표화자의 원핫 벡터를 속성 레이블로 활용하여 음성 변환을 학습한다. 속성 레이블이 학습에 중요한 역할을 담당하며, 안정적인 학습을 한다. 이러한 접근 방식은 원본 화자와 목표화자의 신원을 구분하고 음성 변환 작업을 정확하게 수행하는 데 도움을 준다.

2.4 RawNet3

RawNet3[28]은 딥러닝 기반 모델로, 원시 파형에 대한 화자 인식 성능을 향상하는 데 사용한다. 이 모델은 원시 음성 신호를 직접 입력받아 전처리 과정을 간소화하여, 전처리 과정을 거친 기존 방법과 달리 손실되지 않은 정보를 활용하여 화자 식별 및 화자 검증 성능을 향상한다. 제안된 모델은 Res2Net[32] 기반 모듈과 다중 계층 특징 군집화(Multi-layer feature aggregation)를 사용하여 구성한다.

RawNet3는 RawNet2[33]와 ECAPA-TDNN[34]의 혼합 형태(Hybrid form)인 구조로 구성되어 있으며, 로그(Logarithm) 및 정규화(Normalization) 기능을 포함하고 있다. RawNet2의 학습할 수 있는 필터 뱅크를 사용하여 원시 음성 신호를 시간-주파수 표현으로 변환하는데, 이때 기존의 실수 필터 뱅크 대신 복소수 필터 뱅크를 사용한다. 이러한 구조는 RawNet3에서 신호 처리의 효율성과 성능 개선을 위해 사용되는 중요한 요소이다.

기존 Res2Net의 ECAPA-TDNN 블록을 기반으로 만들어진 AFMS-Res2MP 블록은 RawNet2의 출력값을 요약한다. AFMS-Res2MP 블록은 ECAPA-TDNN과는 다르게 새로운 최대 풀링(max-pooling)과 잔여(residual) 연결을 추가함을 통해 정보를 요약하고 강조한다. 둘째, 기존 RawNet2에서 사용한 squeeze-excitation[35] 대신 특징 지도 재조정(α-feature map rescaling, AFMS) 방식을 적용하여 화자 정보와 관련된 특징값을 강조한다.

이를 거쳐서 나온 특징값은 마지막으로 채널(channel)과 문맥(context)에 종속적인 통계적 풀링(statistic pooling)을 사용하여 시간 축을 줄인다. 이 모델은 지도 학습(Supervised Learning)과 자기 지도 학습(Self-supervised Learning)이 제안되었으며, 본 논문은 지도 학습 방법을 사용한다.

3. 제안된 방법

3.1 네트워크 구조

본 논문은 다양한 화자에 대한 일반화 능력을 갖추고 있으며, 다중 화자에 적용할 수 있는 화자 음성 변환 방법을 제안한다. Fig. 1에서 제안된 음성 변환 모델의 전체 구조를 나타내었다. 먼저, 음성 신호에서 WORLD Vocoder를 사용하여 Spectral envelope(SP), Fundamental frequency(F0), Aperiodic Parameter(AP)를 분리한다. 그리고 사람의 청각 특징을 고려하여 SP를 멜 스케일(Mel scale)로 변환한다. 이

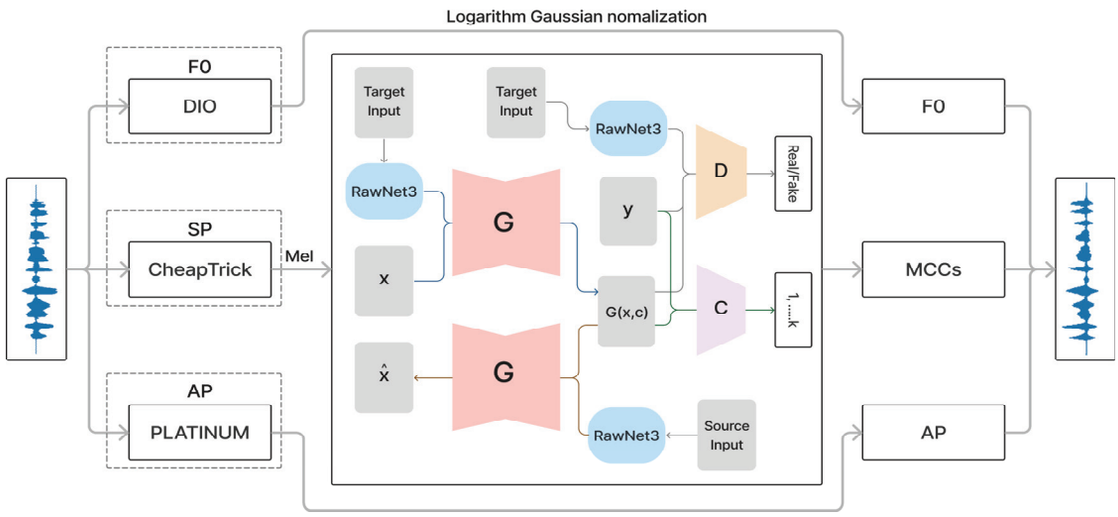


Fig. 1. Architecture Framework of the Proposed Voice Conversion Model

를 통해 음성 신호의 특징을 정확하게 파악하고, 더 적은 차원으로 표현할 수 있다. 변환된 SP로부터는 Mel Cepstral Coefficients(MCCs)를 추출한다. 이렇게 추출된 MCCs를 StarGAN-VC에 입력하여 화자 음성 변환 모델을 학습한다.

본 논문에서는 다양한 화자에 대응하기 위해 기존의 원핫 벡터 대신 RawNet3을 활용하여 추출된 화자의 특징 벡터를 속성 레이블로 사용한다. 이를 통해 다양한 화자의 목소리로 변환할 수 있는 일반화 성능을 향상하고, any-to-any 변환이 가능하여지도록 한다. 제안된 모델에서는 미리 학습된 RawNet3를 동결시켜 사용하며, 필요에 따라 학습할 수 있는 Fully Connected (FC) 계층을 추가한다. 생성기, 판별기, 분류기의 구조는 각각 Fig. 2-4에 나타내어져 있다. 이를 통해 제안된 방법은 다양한 화자 간의 음성 변환 작업을 효과적으로 수행할 수 있다.

이때, “Conv”, “Deconv”, “FC”, “Instance Norm”, “GLU”, “ELU”, “Sigmoid”, “Avgpool”과 “Logsoftmax”는 각각 convolution, deconvolution, fully-connected, instance normalization, gated linear unit, Exponential Linear Unit, sigmoid, average pooling과 log-softmax layer를 나타낸다.

RawNet3에서 추출된 특징 벡터는 생성기와 판별기에서 사용한다. 각 화자의 특성을 추가하는 과정에서 입력 텐서의 차원을 동일하게 맞추기 위해 FC 계층과 Exponential Linear Unit (ELU) 활성화 함수를 사용한다(Fig. 2, Fig. 3).

StarGAN-VC의 성능을 향상하고 학습의 안정성을 높이기 위해 다음과 같은 방법을 적용한다. 첫째, GAN의 적대적 손실함수를 개선할 수 있는 Earth Mover's Distance와 그래디언트 페널티를 도입한다. 둘째, TTUR을 적용한다. 셋째, RNN 대신에 Gated Linear Units(GLU)를 사용한다. GLU 과정에서, 입력의 절반을 나누어 처리하는 방식 대신 전체 입력을 한꺼번에 입력한다.

제안된 모델은 비 병렬적이며 모든 화자에 대한 many-to-many 방법보다 더욱 일반화된 성능을 보여 any-to-any가 가능하다. 즉, 고정된 화자 수에 제한되지 않고 RawNet3를 사용해 추출된 화자의 특징을 활용하여 음성 변환을 수행할 수 있다.

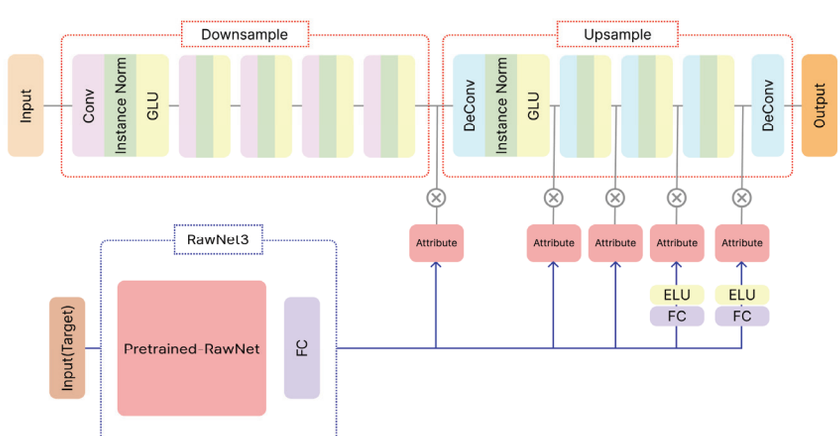


Fig. 2. Generator Architecture Framework of the Proposed Voice Conversion Model

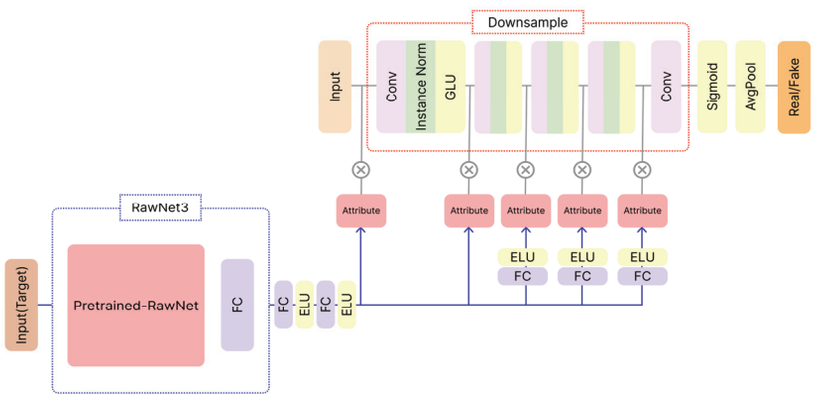


Fig. 3. Discriminator Architecture Framework of the Proposed Voice Conversion Model

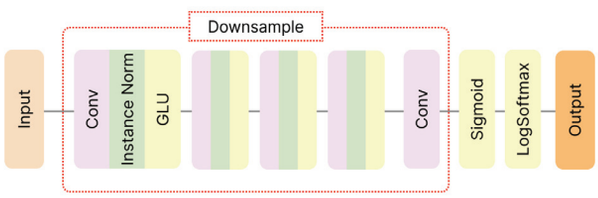


Fig. 4. Classifier Architecture Framework of the Proposed Voice Conversion Model

3.2 손실함수

본 논문은 StarGAN-VC의 손실함수를 활용하여 기존 손실함수의 단점을 보완하여 재구성하였다. 이러한 손실함수들은 제안된 모델의 학습을 돕고, 생성된 음성의 품질과 변환의 일관성을 향상하는 데 이바지한다.

제안된 방법에서는 생성기에서 다음과 같은 손실함수들이 사용한다. 이때, x는 입력으로 사용되는 임의의 원본화자이고 r는 RawNet3으로부터 추출한 목표화자의 특징 벡터이다. $G(x,r)=\hat{y}$ 은 생성기를 거쳐 목표화장이 특성을 지닌 생성된 음성 데이터이다. Equation (2)의 Modified adv 손실함수는 Earth Mover's Distance를 적용한 적대적 손실함수(Adversarial Loss)로써 생성자와 판별자 간의 경쟁을 통해 원본 음성과 변환된 음성 사이의 차이를 줄이는 데 목표로 한다.

$$L_{\mathrm{mod}ified\ adv}(G)= -E_{x\sim\ p(x),r\sim\ p(r)}[(D(G(x,r),r))] \tag{2}$$

Equation (3)의 cls 손실함수는 분류 손실함수(Classification Loss)로써 생성된 음성 데이터가 목표화자 레이블로 알맞게 분류하는지를 판단하는 교차 엔트로피(Cross Entropy)이다. 이 손실함수는 모델의 출력 분포와 실제 값 사이의 교차 엔트로피를 계산하여 사용한다.

$$L_{ds}(G)= -E_{x\sim\ p(x),r\sim\ p(r)}[\log p_r(r\mid G(x,r))] \tag{3}$$

Equation (4)의 cyc 손실함수는 순환 일관성 손실함수(Cycle Consistency Loss)로써 주어진 모델에서 음성 변환을 거치고 다시 원래의 화자로 되돌아왔을 때, 원본 음성과 재구성된 음성 간의 차이를 최소화하는 손실함수이다. 이는 L_1 Loss를 사용하여 계산한다. 이때, r'은 RawNet3으로부터 추출한 원본화자의 특징벡터이다.

$$L_{cyc}(G)= E_{r'\sim\ p(r),x\sim\ p(x\mid r'),r\sim\ p(r)}[\parallel G(G(x,r),r')-x\parallel_1] \tag{4}$$

Equation (5)의 id 손실함수는 정체성 매핑 손실함수(Identity Mapping Loss)로써 생성기에 원본 화자의 음성과 원본 화자의 특징 벡터를 주어 변환하였을 때, 변환된 음성이 원본 음성과 최대한 일치하도록 학습하는 손실함수이다. 이는 L_1 Loss를 사용하여 계산한다.

$$L_{id}(G)= E_{r'\sim\ p(r),x\sim\ p(x\mid r')}[\parallel G(x,r')-x\parallel_1] \tag{5}$$

다음의 Equation (6)은 생성기의 최종 손실함수이다. 이때 람다(λ)는 각 손실함수의 비율을 조절하기 위한 값이다.

$$L_G(G)=L_{\mathrm{mod}ified\ adv}(G)+\lambda_{ds}L_{ds}(G)+\lambda_{cyc}L_{cyc}(G)+\lambda_{id}L_{id}(G) \tag{6}$$

판별기에서는 Equation (7)과 Equation (8)과 같은 손실함수들이 사용되었다. Equation (7)의 Modified adv 손실함수는 WGAN-GP의 판별자의 손실함수로서, 생성기에서 사용되는 Modified adv 손실과 유사한 방식으로 Earth

Mover's Distance를 기반으로 한 적대적 손실함수이다. 이 손실함수는 생성된 음성과 실제 목표화자 음성 간의 거리를 최소화하기 위해 판별자를 학습시킨다. 즉, 생성된 음성과 실제 음성을 잘 구분할 수 있도록 유도하는 역할을 한다.

$$L_{\mathrm{mod}ified\ adv}(D) = \\ -E_{r\sim\ p(r),y\sim\ p(y\mid r)}[D(y,r)] \\ +E_{x\sim\ p(x),r\sim\ p(r)}[D(G(x,r),r)] \tag{7}$$

Equation (8)은 그래디언트 페널티로, 이를 적용하여 GAN의 안정성을 높이고 모드 붕괴 문제를 완화하는 데 사용된다. 그래디언트 페널티는 판별자의 경사를 제약하는 역할을 수행하여, 생성된 음성과 실제 음성 사이의 경사 폭발이나 소멸을 방지한다. 이를 통해 GAN의 학습 안정성을 향상하고 다양한 음성 변환 결과를 얻을 수 있도록 도움을 준다.

$$L_{gp}(D) = E_{\hat{x}}[(\|\nabla_{\hat{x}} D_{src}(\hat{x})\|_2 - 1)^2] \tag{8}$$

다음의 Equation (9)은 판별기의 최종 손실함수이다.

$$L_D(D) = L_{\mathrm{mod}ified\ adv}(D) + \lambda_{gp} L_{gp}(D) \tag{9}$$

Equation (6)은 분류기를 학습시키기 위한 최종 손실함수이다. 이 손실함수는 교차 엔트로피를 기반으로 하며, y인 목표화자의 실제 음성이 정확하게 분류되도록 한다.

$$L_C(C) = -E_{r\sim\ P(r),y\sim\ p(y\mid r)}[\log p_r(r\mid y)] \tag{10}$$

4. 실 험

4.1 데이터셋

1) VCC2018

Voice Conversion Challenge 2018(VCC2018) 데이터셋[36]은 균형 잡힌 화자 구성과 다양한 발화 및 비 병렬 특징이 있다. 이 데이터셋에는 전문적인 영어 원어민 화자 12명의 음성이 담겨있으며, 그중 8명은 원본 화자이고 4명은 목표화자이다. 데이터셋은 남녀 성비가 균등하게 배분되어 있으며, 총 972개의 발화 음성 파일이 포함되어 있다. 이 데이터셋의 샘플링 레이트(Sampling Rate)는 22,050Hz이다.

VCC2018 데이터셋은 음성 변환 알고리즘의 효과적인 개발과 평가를 위해 필요한 다양한 음성적 특징을 제공한다. 데이터셋은 다양한 발화로 구성되어 있어 음성 변환 과정에서 발음과 억양의 다양성을 충분히 고려할 수 있는 중요한 자원이다.

이 데이터셋의 핵심 특징 중 하나는 비 병렬 데이터 구조이다. 이는 병렬 구조에 의존하지 않는 음성 변환 알고리즘의 개발과 평가에 적합하다. 이러한 구조의 데이터셋을 활용하여 개발된 알고리즘은 다양한 실제 상황에서의 적용 가능성을 더욱 효과적으로 평가할 수 있다.

2) ESD

Emotional Speech Database(ESD) 데이터셋[37]은 다양한 언어와 방언, 연령대 및 성별의 화자를 포함하고 있다. 이 데이터셋은 국제적인 음성 변환 연구를 지원하며, 다양한 언어와 문화적 배경을 가진 화자들 사이의 음성 변환에 초점을 맞춘다.

ESD 데이터셋은 16,000Hz의 샘플링 레이트로 영어 원어민 10명과 중국어 원어민 10명이 5가지의 감정(중립, 기쁨, 분노, 슬픔, 놀람)으로 각각 한 명의 한 감정당 350개를 발화 음성이 제공한다. 이를 통해 화자 당 총 1,750개의 발화 음성 파일이 포함되어 있다. 따라서 ESD 데이터셋에는 총 35,000개의 발화 음성 파일이 포함되어 있다. 이 데이터셋은 VCC2018 데이터셋과 마찬가지로 비 병렬 데이터 구조로 되어있다.

4.2 구현 세부 정보

본 논문에서는 VCC2018 데이터셋을 모두 사용하여 음성 변환을 수행하였다. 원본 화자는 8명이며, 목표화자는 4명으로 구성되어 총 32가지의 경우로 음성 변환을 진행하였다. 하지만 RawNet3의 입력 오디오 샘플링 레이트가 16,000Hz로 고정되어 있으므로 샘플링 레이트가 22,050Hz인 VCC2018 데이터셋을 그대로 사용할 수 없다. 따라서 실험을 위해 VCC2018 데이터셋을 16,000Hz로 리샘플링(Resampling)한 후에 진행하였다.

Table 1에서는 실험 설정에 대한 정보를 제공한다. 실험에서는 배치 크기를 4로, Mel 필터 개수를 36으로 설정하였다. 그래디언트 페널티의 람다(λ) 값은 10, Cycle Loss의 람다 값은 10, Classification Loss의 람다 값은 1, 그리고 identification loss의 람다 값은 3으로 가중치를 설정하여 학습을 진행하였다. TTUR를 적용하기 위해 판별기와 생성기의 학습 비율은 3:1로 설정되었다.

생성기의 학습률은 0.0005로 설정하였고, 판별기와 분류기, 그리고 동결되지 않은 부분의 RawNet3 부분의 학습률은 0.0001로 설정하였다. 총 800,000번의 반복 학습을 진행하였으며, 100,000번마다 이전 학습률을 100,000으로 나눈 값을 현재 학습률에서 뺀 값으로 업데이트했다. 실험은 Ubuntu 20.04 운영체제에서 PyTorch 및 GeForce RTX 3090 GPU를 사용하여 수행되었다.

Table 1. Experimental Details

Sampling Rate	16,000 (Hz)
Batch Size	4
Mel	36
그래디언트 페널티의 람다(λ)	10
Cycle Loss의 람다	10
Classification Loss의 람다	1
identitfication loss의 람다	3
생성기의 학습률 (Learning Rate)	0.0005
판별기의 학습률 (Learning Rate)	0.0001
분류기의 학습률 (Learning Rate)	0.0001
RawNet3의 학습률 (Learning Rate)	0.0001
TTUR (판별기:생성기)	3:1

4.3 평가 지표

본 실험에서는 음성 변환 모델의 성능을 평가하기 위해 다음의 네 가지 지표를 사용한다. 첫 번째로 사용한 Mel Cepstral Distortion(MCD)는 변환된 음성과 원본 음성의 Mel Cepstral Coefficients(MCC) 간의 거리를 측정한다. 이 지표는 음성 변환의 정확성을 평가한다. 두 번째 지표는 Mel Spectral Distortion(MSD)로, 변환된 음성의 멜 스펙트럼과 원본 음성 간의 거리 차이를 측정하여 음성 변환의 품질을 평가한다. 세 번째로 사용한 Pitch Conversion Error(PCE)는 변환된 음성의 기본 주파수(fundamental frequency, F0)와 원본 음성의 기본 주파수와의 간의 차이를 측정한다. 이를 통해 음성 변환 과정에서 발생하는 피치 변환의 정확성을 평가할 수 있다. 그리고 네 번째로 사용하여 t-Distributed Stochastic Neighbor Embedding(t-SNE)를 변환된 음성과 원본 음성의 특징 벡터 간의 관계를 시각화한다. 이 시각화를 통해 음성 변환 모델의 성능을 직관적으로 이해할 수 있다. 이러한 지표들을 통해 제안된 음성 변환 모델의 성능을 종합적으로 평가하였다.

5. 실험 결과

본 논문에서는 Fig. 5에서 VCC2018 데이터셋으로부터 추출된 RawNet3의 특징 벡터를 보인다. “SF1”은 원본 음성의 첫 번째 여성 화자를 나타내며, 해당 화자가 발화한 문장들이 정확하게 해당 화자로 군집화되는 것을 실험 결과에서 확인하였다. 이는 제안된 RawNet3 모델이 화자의 특징을 적절하게 추출하여 특징 벡터의 속성값으로 사용할 수 있음을 시각적으로 확인할 수 있다.

기존 방법과 제안된 방법의 성능을 Table 2에서 비교하였다.

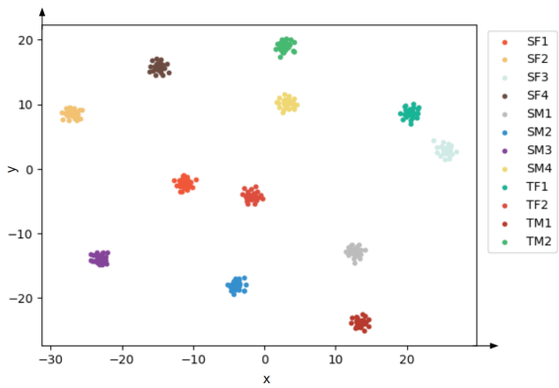


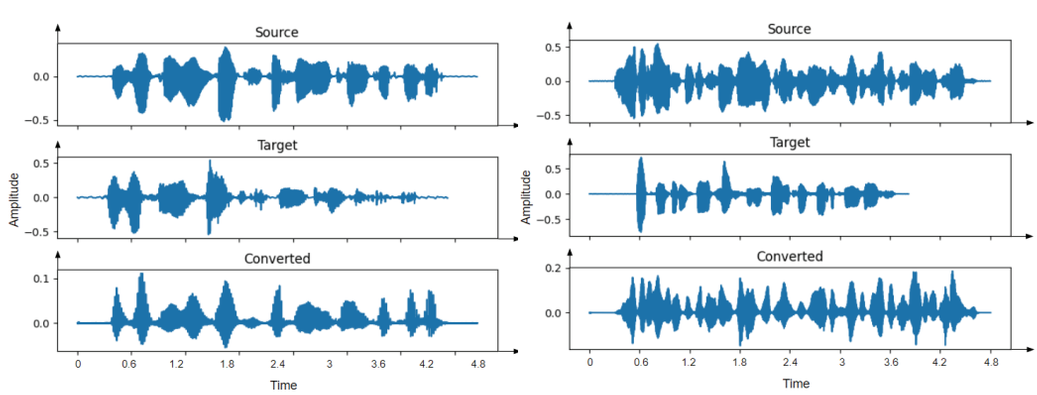
Fig. 5. tSNE of the VCC2018 Dataset after RawNet3

제안된 방법은 StarGAN-VC보다 더 많은 화자를 사용하여 음성 변환을 수행했음에도 불구하고, MCD와 MSD 성능이 각각 6.75와 1.1로 향상되었다. 또한, 학습할 때 이용하지 않았던 ESD 데이터셋을 이용하여 일반화 성능을 확인한 결과, MCD와 MSD가 각각 6.72와 1.09로 높은 성능을 보였다. 이는 학습에 사용되지 않은 ESD 데이터셋으로부터 우수한 성능의 음성 변환을 수행하였으며, 제안된 모델의 일반화 성능이 향상되었음을 확인하였다.

실험 결과를 자세하게 분석하기 위해 원본 화자의 음성, 목표화자의 음성, 그리고 변환된 음성의 파형을 Fig. 6에서 보여주었다. Fig. 6(a)는 병렬 구조로 음성 변환을 수행한 경우를, Fig. 6(b)는 비 병렬 구조로 음성 변환을 수행한 경우를 나타내었다. Fig. 6에서 생성된 음성이 원본 음성의 발화 내용을 유지하면서 목표화자의 음성 특징을 잘 나타내고 있음을 확인하였으며, 비 병렬 구조에서도 잘 변환되었음을 보여주었다.

Table 2. Experimental Results

Model	Dataset	RawNet Attribute	MCD [dB]	MSD [dB]	PCE
StarGAN-VC	VCC2018 (4 people)	-	7.11±.10	2.41±.13	-
StarGAN-VC2	VCC2018 (4 people)	-	6.90±.07	1.89±.03	-
Proposed model	VCC2018 (all)	all	6.75	1.1	0.28
	VCC2018 (all)	one	6.8	1.1	0.28
	ESD (English)	all	6.72	1.09	0.59

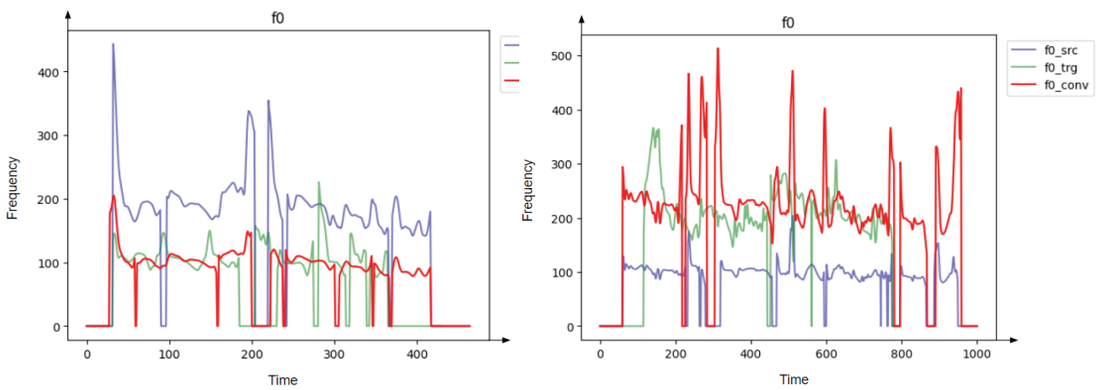


(a) parallel (b) non-parallel

Fig. 6. waveform of Converted Voice Version

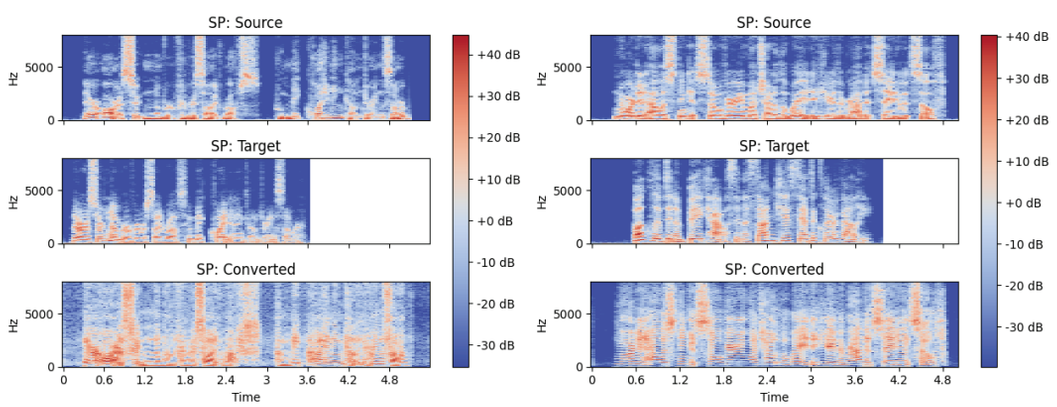
더 자세한 분석을 위해 WORLD Vocoder를 사용하여 F0, AP, SP를 각각 Fig. 7-9에서 각각 보여주었다. Fig. 7에서 F0의 음율과 발화 내용 등 원본 화자의 내용을 유지하면서 목표화자의 목소리와 유사하게 음성이 변환되었음을 확인하였다. Fig. 8의 AP와 Fig. 9의 SP도 병렬 및 비 병렬 구조 여부에는 영향을 받지 않으며, 원본 음성의 특징을 유지하면서 목표화자의 음성 특징을 반영하였음을 보여주었다.

제안된 방법에 따라 변환된 음성이 군집화 정도를 보이는 t-SNE 분포를 Fig. 10-12에서 설명하였다. Fig. 10은 학습에 사용된 VCC2018 학습 데이터셋으로부터 변환된 음성의 t-SNE를 보여주고 있다. Fig. 10(a)는 목표화자를 기준으로 변환된 음성의 t-SNE를 나타내며, Fig. 10(b)는 변환된 음성의 원본 화자 정보를 색상으로 구분하여 나타내었다. 이를 통해 다양한 원본 화자의 목소리가 정확하게 목표화자의 목소리로 변환되었음을 확인할 수 있다.



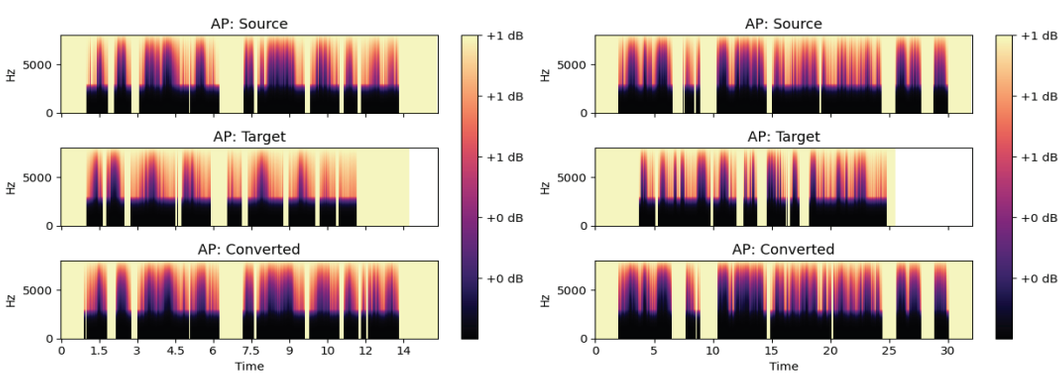
(a) parallel (b) non-parallel

Fig. 7. F0 of Converted Voice Version



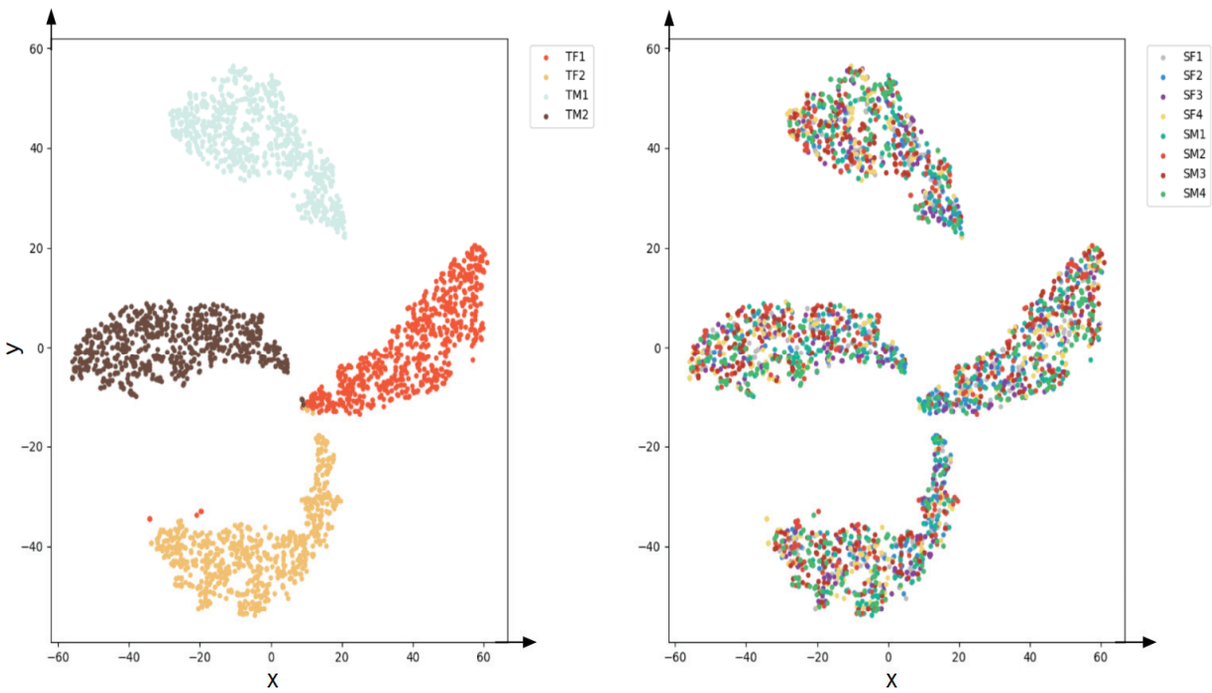
(a) parallel (b) non-parallel

Fig. 8. SP of Converted Voice Version



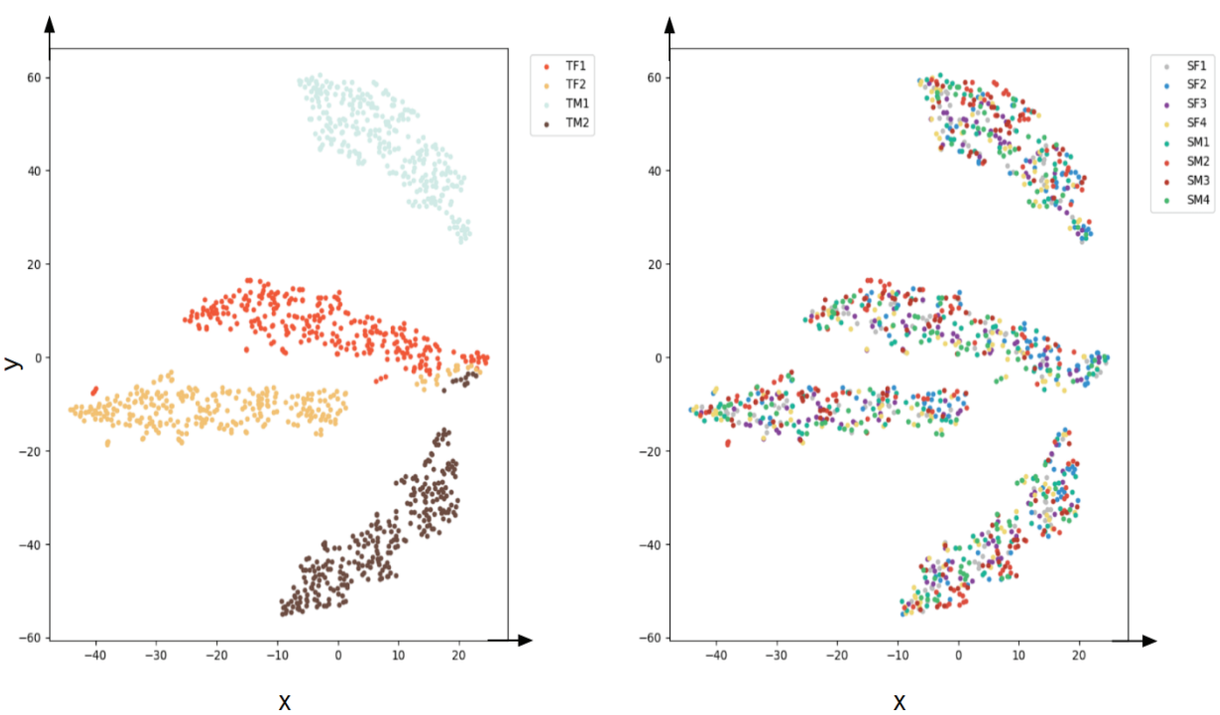
(a) parallel (b) non-parallel

Fig. 9. AP of Converted Voice Version



(a) Color representation by Target (b) Color representation by Source

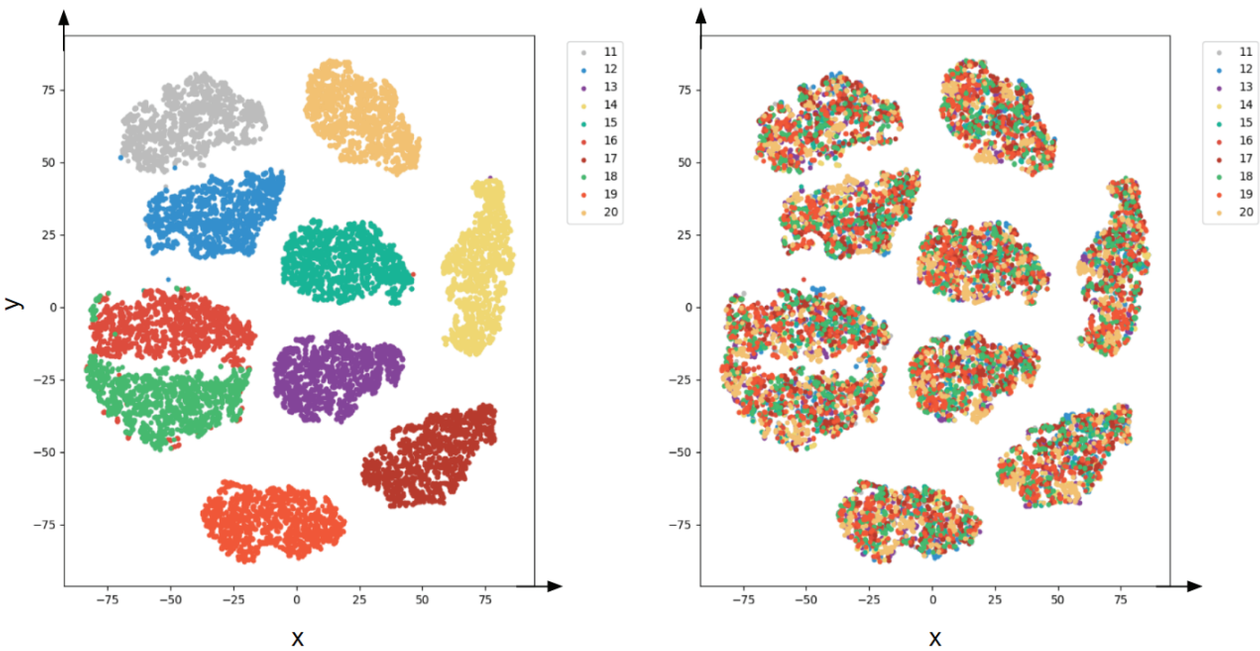
Fig. 10. tSNE of Converted VCC2018 Train Dataset



(a) Color representation by Target (b) Color representation by Source

Fig. 11. tSNE of Converted VCC2018 Test Dataset

또한, 학습에 사용되지 않았던 VCC2018 테스트 데이터셋을 이용하여 변환된 음성의 t-SNE를 Fig. 11에 나타냈다. Fig. 12는 모델의 학습 과정에서 보지 않았던 화자에 대한 ESD 데이터셋으로부터 변환된 음성의 t-SNE를 나타내었다. Fig. 11과 Fig. 12의 t-SNE 분포에서 큰 차이를 보이지 않으며, 이는 화자 변환이 성공적으로 진행되었음을 의미한다.



(a) Color representation by Target (b) Color representation by Source

Fig. 12. tSNE of Converted ESD Test Dataset (Unseen)

6. 결 론

목표화자 정보의 원핫 벡터를 속성으로 이용하는 기존 StarGAN-VC 방법의 제약을 해결하기 위해, 본 논문에서는 RawNet3로부터 추출된 화자의 특징을 속성으로 이용한 StarGAN-VC 기반의 목소리 변환 방법이 제안된다. 이를 통해 발화자에 대한 일반화 성능을 높이고, 다양한 발화자에 대한 any-to-any 음성 변환을 수행할 수 있다. 더욱 안정적인 학습 성능을 위해 Wasserstein-1 Distance 손실함수, 그래디언트 페널티와 TTUR을 도입하였다. 기존 방법과의 비교를 통해 제안된 방법의 개선된 목소리 변환 성능을 확인하였다.

References

[1] Y. Wang, P. Getreuer, T. Hughes, R. F. Lyon, and R. A. Saurous, "Trainable frontend for robust and far-field keyword spotting," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.5670-5674, 2017.

[2] W. Fan, X. Xu, B. Cai, and X. Xing, "ISNet: Individual standardization network for speech emotion recognition," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol.30, pp.1803-1814, 2022.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3156-3164, 2015.

[4] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang, "Hierarchical recurrent neural encoder for video representation with application to captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1029-1038, 2016.

[5] D. Amodei et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, pp.173-182, 2016.

[6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2Vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, Vol.33, pp.12449-12460, 2020.

[7] J. Shen et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.4779-4783, 2018.

[8] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "FastSpeech: Fast, Robust and Controllable Text to Speech," *Advances in Neural Information Processing Systems*, Vol. 32, 2019.

[9] J. Yeung and G. Bae, Forever young, beautiful and scandal-free: The rise of South Korea's virtual influencers [Internet], https://edition.cnn.com/style/article/south-korea-virtual-influencers-beauty-social-media-intl-hnk-dst/index.html

[10] J. Zong, C. Lee, A. Lundgard, J. W. Jang, D. Hajas, and A. Satyanarayan, "Rich screen reader experiences for accessible data visualization," *Computer Graphics Forum*, Vol.41, No.3, 2023.

[11] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*, pp.5210-5219, 2019.

[12] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.3893-3896, 2009.

[13] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol.29, pp.132-157, 2020.

[14] D. A. Reynolds, "Gaussian mixture models," *Encyclopedia of biometrics*, Vol.741, pp.659-663, 2009.

[15] S. Mobin and J. Bruna, "Voice conversion using convolutional neural networks," arXiv preprint, 2016. [Internet], https://arxiv.org/abs/1610.08927

[16] J. Lai, B. Chen, T. Tan, S. Tong, and K. Yu, "Phone-aware LSTM-RNN for voice conversion," in *2016 IEEE 13th International Conference on Signal Processing*, pp.177-182, 2016.

[17] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, Vol.35, No.1, pp.53-65, 2018.

[18] M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv preprint, 2014. [Internet], https://arxiv.org/abs/1411.1784

[19] 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*, pp.2223-2232, 2017.

[20] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," *Advances in Neural Information Processing Systems*, Vol.30, 2017.

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, Vol.20, 2017.

[22] 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)*, pp.266-273, 2018.

[23] Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.8789-8797, 2018.

[24] M. S. Al-Radhi, T. G. Csapó, and G. Németh, "Parallel voice conversion based on a continuous sinusoidal model," in *2019 International Conference on Speech Technology and Human-Computer Dialogue*, pp.1-6, 2019.

[25] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.5274-5278, 2018.

[26] W. C. Huang, T. Hayashi, Y. C. Wu., H. Kameoka, and T. Toda, "Pretraining techniques for sequence-to-sequence voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol.29, pp.745-755, 2021.

[27] S. Lee, B. Ko, K. Lee, I. C. Yoo, and D. Yook, "Many-to-many voice conversion using conditional cycle-consistent adversarial networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.6279-6283, 2020.

[28] J. W. Jung, Y. J. Kim, H.S. Heo, B. J. Lee, Y. Kwon, and J. S. Chung, "Pushing the limits of raw waveform speaker recognition." in *Proceedings of Interspeec*, pp.2228-2232, 2022.

[29] M. Morris, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications." *IEICE TRANSACTIONS on Information and Systems*, Vol.E99-D, No.7, pp.1877-1884, 2016.

[30] E. O. Brigham and R. E. Morrow, "The fast Fourier transform," IEEE Spectrum, Vol.4, No.12, pp.63-70, 1967.

[31] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks," in *Proceedings of the 34th International Conference on Machine Learning*, Vol.70, pp.214-223, 2017.

[32] S. H. Gao, M. M. Cheng, K. Zhao, and X. W. Hu, "Res2Net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.43, No.2, pp.652-662, 2019.

[33] J. W. Park, S. B. Kim, H. J. Shim, J. H. Kim, and H. J. Yu, "Improved RawNet with Feature Map Scaling for Text-independent Speaker Verification using Raw Waveforms," in *Proceedings of Interspeech*, pp.1496-1500, 2020.

[34] B. Desplanques, J. Thienpondt, K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proceedings of Interspeech*, pp.3830-3834, 2020.

[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.7132-7141, 2018.

[36] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proceedings of The Speaker and Language Recognition Workshop (Odyssey 2018)*, pp.195-202, 2018.

[37] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.920-924, 2021.

박 보 경

https://orcid.org/0009-0003-8271-8218
e-mail : pym256@naver.com
2021년 동덕여자대학교 정보통계학부 (학사)
2021년 ~ 현 재 중앙대학교 AI학과 석사과정
관심분야 : Deep Learning, Audio Recognition, Computer Vision

박 소 민

https://orcid.org/0009-0002-5864-6501
e-mail : caqta0323@naver.com
2021년 중앙대학교 융합공학부(학사)
2021년 ~ 현 재 중앙대학교 컴퓨터공학과 석사과정
관심분야 : Deep Learning, Audio, Speech Emotion Recognition, Computer Vision

홍 현 기

https://orcid.org/0000-0002-0815-9492
e-mail : honghk@cau.ac.kr
1998년 중앙대학교 전자공학과(박사)
1999년 서울대학교 자동제어특화연구센터 연구원
2002년 Univ. of Colorado, Post-doc.
2000년 ~ 현 재 중앙대학교 첨단영상대학원, 소프트웨어대학 소프트웨어학부 교수
관심분야 : Deep Learning, Speech Conversion, Speech Emotion Recognition, Computer Vision