KIPS Trans. Softw. and Data Eng.
Vol.10, No.1 pp.9~18
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2021.10.1.9

Improving Fidelity of Synthesized Voices Generated by Using GANs

Moon-Ki Back† · Seung-Won Yoon†† · Sang-Baek Lee†† · Kyu-Chul Lee†††

ABSTRACT

Although Generative Adversarial Networks (GANs) have gained great popularity in computer vision and related fields, generating audio signals independently has yet to be presented. Unlike images, an audio signal is a sampled signal consisting of discrete samples, so it is not easy to learn the signals using CNN architectures, which is widely used in image generation tasks. In order to overcome this difficulty, GAN researchers proposed a strategy of applying time-frequency representations of audio to existing image-generating GANs. Following this strategy, we propose an improved method for increasing the fidelity of synthesized audio signals generated by using GANs. Our method is demonstrated on a public speech dataset, and evaluated by Fréchet Inception Distance (FID). When employing our method, the FID showed 10.504, but 11.973 as for the existing state of the art method (lower FID indicates better fidelity).

Keywords : Generative Adversarial Networks, Fréchet Inception Distance, Fidelity Improvement, Synthesized Voice

GAN으로 합성한 음성의 충실도 향상

백 문 기† · 윤 승 원†† · 이 상 백†† · 이 규 철†††

요 약

생성적 적대 신경망(Generative Adversarial Networks, GANs)은 컴퓨터 비전 분야와 관련 분야에서 큰 인기를 얻었으나, 아직까지는 오디오 신호를 직접적으로 생성하는 GAN이 제시되지 못했다. 오디오 신호는 이미지와 다르게 이산 값으로 구성된 샘플링된 신호이므로, 이미지 생성에 널리 사용되는 CNN 구조로 학습하기 어렵다. 이러한 제약을 해결하고자, 최근 GAN 연구자들은 오디오 신호의 시간-주파수 표현을 기존 이미지 생성 GAN에 적용하는 전략을 제안했다. 본 논문은 이 전략을 따르면서 GAN을 사용해 생성된 오디오 신호의 충실도를 높이기 위한 개선된 방법을 제안한다. 본 방법은 공개된 스피치 데이터세트를 사용해 검증했으며, 프레쳇 인셉션 거리(Fréchet Inception Distance, FID)를 사용해 평가했다. 기존의 최신(state-of-the-art) 방법은 11.973의 FID를, 본 연구에서 제안하는 방법은 10.504의 FID를 보였다(FID가 낮을수록 충실도는 높다).

키워드 : 생성적 적대 신경망, 프레쳇 인셉션 거리, 충실도 개선, 합성된 음성

1. 서 론

일반적으로 판별 모델(discriminative model)의 분류 성능은 주어진 학습 데이터의 양에 비례하며, 모델의 복잡도에 비해 적은 양의 학습 데이터를 사용하면 과적합(overfitting) 문제가 쉽게 발생하는 단점이 있다. 데이터 증강(data augmentation)은 과적합 문제를 개선하기 위해 주어진 학습 데이터를 양적으로 보충하는 기법으로 딥러닝 연구에서 흔하게 사용한다. 대표적으로 주어진 데이터를 크기조정(scaling) 하거나, 회전(rotation) 및 반전(reflection)하여 학습 데이터의 양을 늘리는 방식은 판별 모델을 개발하는 연구에서 흔하게 찾아볼 수 있으며[1], 이러한 전통적인 데이터 증강 기법은 여러 딥러닝 챌린지에서 효과가 입증되었다[2-4]. 그러나 기하학적 변환을 가한 증강된 데이터가 실제 데이터(모집단)에 포함되는지 검증하는 절차가 없다는 단점이 있다. 이는 증강된 데이터에 현실과 동떨어진(unrealistic) 데이터가 다수 포함될 수 있음을 의미하며, 이를 학습한 판별 모델은 실제 세계에서 좋은 성능을 보이기 어려울 것이다. 즉 전통적인 데이터 증강은 과적합을 개선하는 효과가 있지만, 모델의 일반화(generalization)성능을 개선시키는데 한계가 있다.

최근에는 생성적 적대 신경망(Generative Adversarial Network, GAN)[4]을 사용하여 전통적인 데이터 증강의 한계점을 개선하려는 연구가 대두되고 있다. GAN은 생성 모델(generative model)의 한 종류로, 주어진 학습 데이터의 확

※ 이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2019R1A2C1011567).
† 준 회 원 : 충남대학교 컴퓨터융합학부 석·박사통합과정
†† 비 회 원 : 충남대학교 컴퓨터융합학부 석·박사통합과정
††† 정 회 원 : 충남대학교 컴퓨터융합학부 교수
Manuscript Received : August 20, 2020
Accepted : October 18, 2020
* Corresponding Author : Kyu-Chul Lee(kclee@cnu.ac.kr)

률 분포를 학습하여 학습된 분포로부터 새로운 데이터를 생성할 수 있다. 그래서 전통적인 데이터 증강과 다르게 매우 자연스러운(realistic) 데이터를 생성할 수 있다. 대표적으로 [6,7]은 GAN을 이용해 가상의 이미지를 생성한 연구로, 사람조차 실제와 가상의 이미지를 구분하기 어려울 정도이다. 특히 정답 데이터(ground truth)를 입수하는데 큰 비용이 요구되는 의료 분야의 데이터 증강 기법으로서 활용가치가 높다[8].

GAN은 컴퓨터 비전 분야를 구심점으로 비약적인 발전을 거듭하고 있지만, 아직까지는 GAN을 사용해 소리 데이터를 직접적으로(independently), 그리고 안정적으로 생성할 수 있는 참조 모델은 제시되지 못한 상황이다. 파형(waveform)과 같은 소리 데이터는 이미지와 다르게 이산(discrete) 값이 시간에 따라 나열된 데이터로, 시간적 특징(temporal feature)이 강하게 포함되어 있다. 그래서 이미지를 학습하는 생성 모델에 주로 사용되어온 합성곱 신경망(Convolutional Neural Networks, CNN) 구조로는 소리 데이터를 학습시키기 어렵다. 그래서 GAN 연구자들은 시간-주파수 표현(time-frequency representation)을 사용하여 소리 데이터를 학습하는 전략을 제안하였다. 스펙트로그램은 대표적인 시간-주파수 표현으로, 이미지와 같이 행렬로 표현되기 때문에 시각화할 수 있으며, 안정적인 학습이 가능한 기존의 이미지 생성 GAN을 재사용하여 학습이 가능하다.

본 논문은 소리 데이터의 범위를 음성으로 한정하고, 최고 성능(state-of-the-art)을 보인 기존 GAN 기반 소리 생성 연구를 분석하여 충실도(fidelity) 측면의 개선 방법을 제안한다. 충실도는 생성 모델의 대표적인 평가 지표로, GAN이 생성한 데이터가 얼마나 현실적인지를 정량적으로 나타낸다. 본 연구에서는 기존 연구에서 다루지 않은 새로운 특징을 학습에 사용하여 GAN이 생성한 소리 데이터의 충실도를 높이는 것이 주요 기여이다.

본 논문의 구성은 다음과 같다. 2장에서는 GAN을 이용해 소리 데이터를 생성하는 최신 연구를 소개하고, 3장에서는 본 연구에서 사용하는 시간-주파수 표현을 상세하게 다루면서, GAN이 생성한 소리 데이터의 충실도를 개선하기 위한 방법을 제안한다. 4장에서는 개선된 방법을 GAN에 적용하여 생성된 소리 데이터의 충실도가 얼마나 개선되었는지 정량적으로 평가하며, 마지막으로 5장에서는 결론 및 향후 연구에 대하여 논한다.

2. 관련 연구

GAN은 가상의 얼굴 이미지를 생성하는 시연과 함께 컴퓨터 비전 분야를 중심으로 큰 주목을 받았다. 그리고 CNN 기반의 DCGAN[9]이 등장하면서 현재까지 이미지 생성 분야의 비약적인 발전이 이루어졌다. 현재의 GAN은 현실에 있을법한 새로운 가상의 이미지를 생성할 수 있어서, 주로 학습 데이터 입수에 시간과 비용이 많이 소모되는 분야에 데이터 증강 기법으로서 활용되고 있다.

GAN 기반의 이미지 생성 연구와 다르게, GAN을 기반으로 소리 데이터를 생성하는 연구는 비교적 최근에 등장했다. WaveGAN[10]은 처음으로 GAN을 사용해 소리 데이터를 생성한 연구로 2018년에 발표되었다. WaveGAN 저자들은 기존 DCGAN의 구조를 변형하여 파형을 학습하고 생성했다: 그레이스케일(gray scale) 이미지를 학습하는 DCGAN의 전체적인 쉐입(shape)을 1차원으로 변형함으로써 16,384차원의 벡터를 생성하는 GAN을 제안 하였다. 특히 WaveGAN의 저자들은 생성기(generator)의 전치 컨볼루션(transposed convolution)에서 발생하는 체커보드 형태의 아티팩트(checkerboard artifacts)[11]가 GAN의 학습을 불안정하게 하는 요인이라고 분석하였다. 구체적으로 체커보드 아티팩트는 프레임 단위로 처리되는 업샘플링(up-sampling) 과정에서 발생하는 일종의 패턴으로, 판별기(discriminator)는 체커보드 아티팩트의 유무만으로 생성기가 생성한 데이터를 쉽게 판별할 수 있다. 따라서 판별기가 생성기를 압도하여 학습이 진행되지 않는 문제가 발생한다. 이 문제를 개선하기 위해 저자들은 판별기의 특징 맵(feature map)을 임의 샘플만큼 순환시킴으로써 판별기가 체커보드 아티팩트를 패턴으로 인식하지 않도록 유도하는 페이즈 셔플(Phase Shuffle)을 제안하였다. 이 페이즈 셔플은 실험을 통해 학습 과정의 안정성을 개선할 수 있음을 보였다. WaveGAN은 GAN을 사용해 소리 데이터를 생성한 최초의 연구이기 때문에, 이후 등장하는 관련 연구들에서 성능평가의 베이스 라인으로 설정하고 있다.

Google AI에서는 2019년에 GAN을 이용해 악기 소리를 생성하는 GANSynth[12]를 발표하였다. GANSynth의 저자들은 먼저 WaveGAN에 모드 붕괴(mode collapse) 문제가 있음을 지적했다. 모드 붕괴가 발생한 GAN은 주어진 학습 데이터의 전반적인 분포를 학습하지 못하여 특정한 범주에 편향된 데이터를 생성하게 된다. 즉, 생성된 데이터의 다양성(diversity)이 매우 낮다. GANSynth의 저자들은 파형을 직접 학습하는 방식이 아닌, 이미지 형태로 시각화 가능한 스펙트로그램을 생성하는 GAN을 제안했다. 스펙트로그램은 그리핀-림(Griffin-Lim)과 같은 알고리즘을 통해 파형으로 변환가능하다. 이처럼 간접적으로 소리 데이터를 생성하는 방식은 WaveGAN의 저자들도 SpecGAN이라고 이름 붙인 실험을 통해 가능한 전략임을 보이긴 했으나, SpecGAN이 생성한 스펙트로그램을 파형으로 변환하여 사람의 청각으로 평가했을 때 충실도가 낮았다고 한다. 즉, 사람의 청각은 WaveGAN 방식으로 생성한 파형을 더욱 자연스럽다고 인식했다. 이에 대하여 GANSynth 저자들은 SpecGAN이 고해상도 스펙트로그램을 학습하지 못하는 DCGAN을 기반으로 설계되어 있기 때문에 충실도를 높이기 어렵다고 분석했다. 그래서 고해상도 이미지를 안정적으로 생성한 PGGAN[13]의 구조를 차용하여 고해상도 스펙트로그램을 생성하는 GAN을 제안했다. 특히, GANSynth 저자들은 생성기의 전치 컨볼루션과 스펙트로그램을 학습시키기 위해 사용하는 푸리에 변환(Fourier Transform)이 프레임 단위로 연산되면서

발생하는 문제점에 주목했다. 이 문제점은 프레임 단위로 연산되는 과정으로 인하여 파형에 전반적으로 나타나는 위상(phase) 정보가 스펙트로그램에 반영되지 못하는 현상으로, 파형의 위상 정보를 학습하지 못한 GAN은 생성된 소리 데이터의 충실도를 높이는데 한계가 있다고 보았다. 그래서 이 위상 정보를 GAN이 학습할 수 있도록, 주어진 파형에 일정하게 나타나는 위상 정보를 표현한 순시 주파수(Instantaneous Frequency)를 학습에 사용하는 전략을 제안했다. 그 결과 GANSynth는 WaveGAN 보다 생성된 소리 데이터의 충실도가 높았으며, 생성된 소리 데이터의 다양성도 크게 개선되었다.

하지만 GANSynth가 다양한 범주의 소리 데이터를 사용하여 성능평가가 이루어지진 않았다. GANSynth가 학습에 사용한 소리 데이터는 NSynth Dataset[14]으로, 여러 종류의 악기에 대하여 특정한 음높이(pitch)를 짧게 연주한 소리이다. 이 악기 소리들은 여러 악기 소리가 동시에 연주된 믹스쳐(mixture) 형태가 아니고, 정현파(sine wave)와 같이 비교적 단순한 주파수로 구성되어있다. 그래서 순시 주파수를 사용해 본래 파형의 위상 정보를 잘 표현할 수 있다. 그러나 음성이나 자연에서 발생하는 소리 들은 여러 주파수의 파형이 혼합된 믹스쳐 형태에 가깝다. 그래서 음성 인식과 같은 딥러닝 분야에 데이터 증강으로서 GANSynth를 활용하기 위해서는 보다 다양한 범주의 소리 데이터에 대한 성능평가가 제공될 필요가 있다.

3. 시간-주파수 표현 및 HPSS

소리는 샘플링(sampling) 및 양자화(quantization)를 통해 이산 값들로 구성된 디지털 신호로 표현된다. 이 디지털 신호는 시간에 대한 함수인 그래프 형태의 파형이므로 푸리에 변환을 이용해 주파수에 대한 함수인 스펙트럼(spectrum)으로 표현함으로써 해당 파형의 주파수 성분을 확인할 수 있다. 주파수 성분이 다르다는 것은 음색(timbre)이 다름을 의미하는 것으로, 서로 다른 소리를 판별하는 근거로 사용된다. 그러나 자연적으로 발생하는 소리들은 시간에 따라 여러 음색이 나타난다. 그래서 주어진 파형이 시간에 따라 어떠한 주파수 성분으로 변화되는지 분석하기 위해 스펙트로그램과 같은 시간-주파수 표현으로 변환하여 분석하는 방법이 널리 사용되고 있다. 본 논문에서는 스펙트로그램을 Equation (1)과 같은 이산-시간 푸리에 변환(Discrete-Time Fourier Transform, DTFT)을 사용하여 n차원 실 벡터(real vector) 형태의 파형을 복소 행렬(complex matrix)로 변환한 X로 한정한다.

$$X(\Omega) = F[x[n]] = \sum_{k=-\infty}^{\infty} x[n]e^{-j\Omega n} \quad (1)$$

스펙트로그램은 행렬이므로 공간적(spatial) 특징을 잘 학습하는 CNN을 사용해 학습할 수 있다. 이러한 접근은 판별 모델을 개발하는 연구에서 흔하게 사용하는 접근이며, [3]과 같이 소리 데이터를 분류하는 딥러닝 챌린지에서도 쉽게 찾아볼 수 있는 접근이다. 그러나 CNN은 이미지와 같은 실 행렬을 주로 학습해왔기 때문에, 복소 행렬의 스펙트로그램을 학습하기 위해서는 추가적인 변환과정이 요구된다. 대표적으로 복소 행렬에 절댓값을 취하여 실 행렬로 변환하거나, 복소 영역(complex domain)에서 학습이 이루어지는 인공 신경망을 고안하여 학습하는 방법[15]이 있다. 본 논문은 실 영역(real domain)에서 소리 데이터를 다루는 기존 GAN 기반 소리 생성 연구를 따르는 연구로, 복소수 변수로 구성된 (complex-valued) 인공 신경망은 고려하지 않는다.

3.1 시간-주파수 표현

본 논문에서는 시간-주파수 표현인 로그-멜 규모 스펙트로그램(Log-mel Magnitude Spectrogram)과 멜 순시 주파수(Mel Instantaneous Frequency)를 제안하는 GAN의 주요 학습 데이터로 사용한다. Fig. 1은 주어진 파형으로부터 이 두가지 시간-주파수 표현을 획득하는 과정을 담은 흐름도이며, 과정에 대한 설명은 다음과 같다.

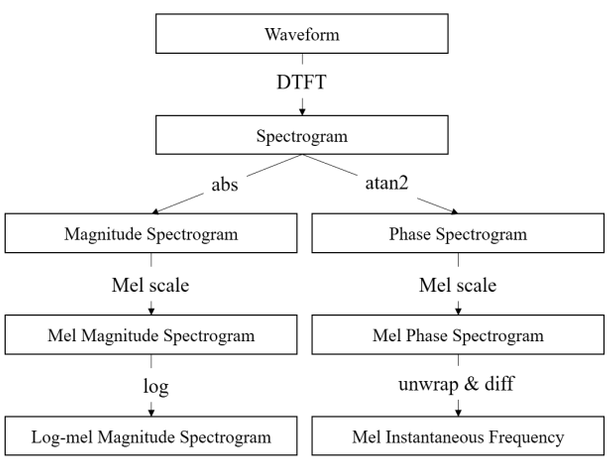


Fig. 1. Flow Diagram for Decomposing a Waveform into Time-frequency Representations

a) DTFT의 복소 지수(complex exponential) 부는 오일러 공식(Euler's formula)을 사용하여 복소 평면에 표현된다. 따라서 복소 행렬 X에 요소 별(element-wise) 절댓값을 취하여 크기를, 역탄젠트(arctangent)를 취하여 편각(argument)을 계산한다.
b) 사람의 청각이 주파수를 인식하는 단위로 표현하기 위하여, 선형 스케일(linear scale)로 이루어진 주파수 축을 멜 스케일(mel scale)로 변경한다.
c) 멜 규모 스펙트로그램(Mel Magnitude Spectrogram)을 로그 스케일로 변경함으로써 사람의 청각이 인식하는 에너지(소리의 크기) 단위로 변경하며, 멜 위상 스펙트로그램(Mel Phase Spectrogram)은 Equation (2)에 의해 멜 스케일의 순시 주파수를 산출한다.

$$\omega_i(t) = \frac{1}{2\pi}\frac{d\theta(t)}{dt} \tag{2}$$

멜 순시 주파수는 GANSynth의 저자들이 GAN을 학습시키는데 사용할 것을 제안하였다. 멜 순시 주파수는 프레임 단위로 처리되는 DTFT 과정에서 위상이 어긋나는 문제를 보완하기 위하여, 2π 범위 내에서 누적하여 더해진(unwrapped) 위상인 $\theta(t)$의 미분을 계산함으로써 본래 파형에 일정하게 나타나는 위상을 실 행렬에 표현한다. GANSynth의 저자들은 논문에 언급하지 않았지만, 멜 규모 스펙트로그램과 멜 순시 주파수는 동일한 파형으로부터 산출되었기 때문에 수학적으로 상관관계를 유도할 수 있으나, 실 영역에서 일반화될 수 없는 이 상관관계를 GAN이 학습을 통해 근사하도록 함으로써 GAN이 생성한 스펙트로그램의 충실도를 개선한다고 해석할 수 있다.

3.2 시간-주파수 표현을 파형으로 변환

생성기는 3.1절의 로그-멜 규모 스펙트로그램과 멜 순시 주파수를 생성하고, 이를 판별기가 판별하는 방식으로 학습이 이루어진다(본 논문에서 제안하는 GAN의 구조는 4장에서 자세하게 다룬다). 만약 판별 모델을 개발하는 연구라면, 주어진 파형을 잘 분류하는 모델을 도출하는 것이 목표이므로 본 절에서 다루는 변환 과정은 불필요할 것이다. 반면 생성 모델은 주어진 파형과 유사한 새로운 파형을 생성하는 것이 목표이므로, GAN을 통해 생성한 시간-주파수 표현을 파형으로 변환할 수 있어야 한다. Fig. 2는 생성기가 생성한 실 행렬 형태의 로그-멜 규모 스펙트로그램과 멜 순시 주파수를 복소 행렬의 스펙트로그램으로 변환하는 과정을 담고 있다. 그리고 이 스펙트로그램은 Equation (3)과 같은 이산-시간 역 푸리에 변환(Inverse DTFT, IDTFT)을 통해 파형으로 복원된다. 이 변환 과정은 다음 설명과 같다.

$$x[n] = F^{-1}[X(\Omega)] = \frac{1}{2\pi}\int_{2\pi} X(\Omega)e^{j\Omega n}d\Omega \tag{3}$$

a) 생성된 로그-멜 규모 스펙트로그램은 요소 별 지수 값을 계산하여 멜 규모 스펙트로그램으로, 생성된 멜 순시 주파수는 스칼라 배(scalar multiplication)하고 누적 합을 계산함으로써 멜 위상 스펙트로그램으로 변환한다.

b) 멜 규모 스펙트로그램은 정방 행렬이 아닐 수 있으므로 무어-펜로즈 유사역행렬(Moore-Penrose pseudoinverse)을 통해 선형 스케일로 변경하며, 위상 스펙트로그램도 같은 과정으로 선형 스케일로 변경한다.

c) 위상 스펙트로그램은 오일러 공식을 통해 복소 행렬로 변환하고, 동일한 쉐입의 규모 스펙트로그램과 요소 별 곱셈을 통해 복소 행렬의 스펙트로그램으로 변환한다. 마지막으로 스펙트로그램은 IDTFT를 통해 n차원 실 벡터인 파형으로 복원한다.

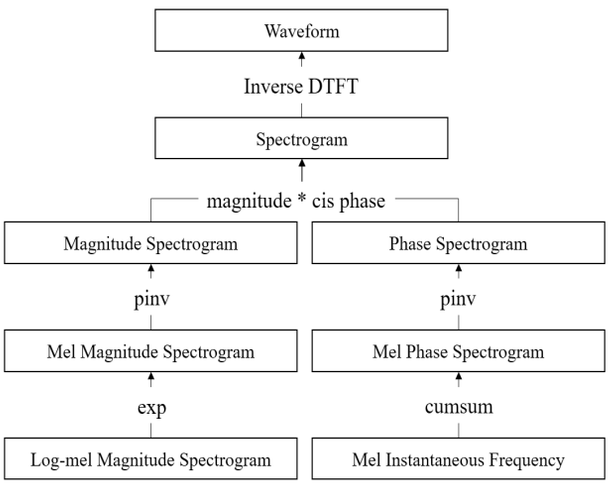


Fig. 2. Flow Diagram for Reconstructing a Waveform from Time-frequency Representations

3.3 HPSS(Harmonic Percussive Source Separation)

HPSS는 주어진 파형을 고조파(harmonic) 성분과 퍼커시브(percussive) 성분으로 분해하는 기법이다. 고조파 성분은 특정한 주파수가 일정하게 유지될 때 나타나고, 퍼커시브 성분은 파형이 반복적으로 등장할 때 나타난다. 이러한 특성은 여러 악기 소리가 섞인 음악에서 쉽게 드러나는데, 음의 높낮이가 다른 소리는 고조파 성분으로, 드럼과 같이 주기를 가지고 반복 등장하는 소리는 퍼커시브 성분으로 나타난다. 두 성분을 분해하는 원리는 고조파와 퍼커시브 성분이 스펙트로그램 상에 다른 모양으로 나타나는 시각적 특징에 기인한다. 고조파 성분은 스펙트로그램의 시간 방향을 기준으로 수평선으로 나타나고 퍼커시브 성분은 수직선으로 구분된다. 이러한 특징 차이는 이미지 처리 분야에서 널리 사용되는 중간 값 필터링(median filtering)과 바이너리 마스킹(binary masking)을 사용하여 분해한다[16]. 그래서 분해된 두 가지 스펙트로그램의 요소 별 합은 원본 스펙트로그램과 같으며, 분해된 두 스펙트로그램을 IDTFT를 통해 각각 고조파 파형과 퍼커시브 파형으로 변환하여도 두 파형의 요소 별 합은 본래의 파형과 같다.

Fig. 3은 악기 소리와 음성이 스펙트로그램 상에서 어떠한 시각적 차이가 나타나는지를 담고 있는 예시이다. 각 스펙트로그램은 악기 소리를 연주한 NSynth Dataset과 여러 사람들이 짧은 단어를 발음한 Speech Commands Dataset[17]으로부터 파형을 추출하여 변환하였다. 이 예시에서 보이는 것처럼, 악기 소리는 스펙트로그램 상에 고조파(수평선) 성분이 주로 나타나며, 음성은 고조파와 퍼커시브(수직선) 성분이 고르게 나타나는 차이점이 있다.

본 논문에서는 Fig. 3과 같이 악기 소리와 음성에 고조파와 퍼커시브 성분이 다르게 분포한다는 차이점으로부터. 아래와 같은 가설을 세우고 교차 상관관계(cross-correlation)를 측정해 보았다.

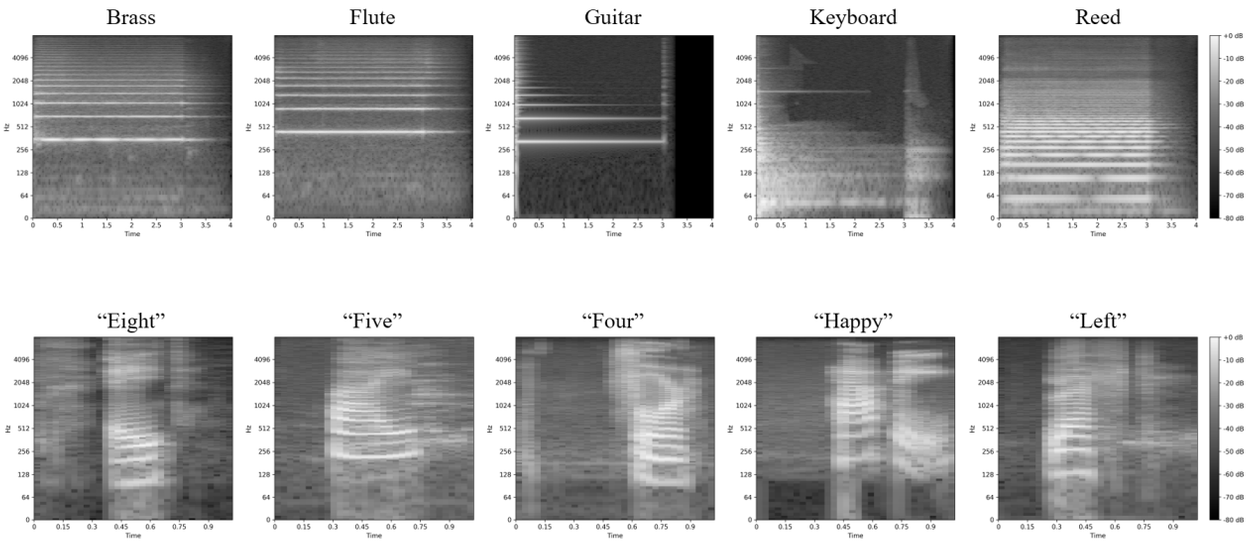


Fig. 3. Comparison of Spectrograms: (top) Spectrograms of Different Kinds of Musical Instruments, (bottom) Spectrograms of Short Words Pronounced by Different Speakers. Horizontal Lines Show the Harmonic Sources and Vertical Lines Represent Percussive Sources

가설: HPSS를 이용해 파형을 고조파 및 퍼커시브 파형으로 분해하고, 두 파형에 대하여 시간-주파수 표현으로 변환하여 획득한 네 종류의 시간-주파수 표현을 다시 파형으로 복원 했을 때, HPSS를 적용하지 않았을 때 보다 본래 파형과 더 유사할 것이다. 즉, HPSS를 사용하면 복소 영역의 스펙트로그램이 실 영역의 시간-주파수 표현으로 변환되며 손실되는 정보가 더 적을 것이다.

위 가설을 확인하기 위하여 앞서 언급한 NSynth Dataset과 Speech Commands Dataset을 대상으로 가설을 적용하여 교차 상관관계를 측정했다.

먼저 GANSynth 저자들이 학습 데이터로 사용한 NSynth Dataset을 대상으로 원본 파형과 실 행렬의 시간-주파수 표현으로부터 복원된 파형 사이의 교차 상관관계를 측정하였다. NSynth Dataset에서 테스트 셋으로 레이블링된 4,096개의 파형에 대하여 교차 상관관계를 측정한 결과는 Fig. 4의 히스토그램과 같다. 각 히스토그램에 표기된 약자의 의미는 아래와 같으며, 표기된 수치는 측정된 교차 상관관계에 대한 평균±표준편차이다(교차 상관관계가 1이면 복원된 파형이 원본과 같음을 의미한다).

a) vs. O: 원본 파형과 원본 파형 사이의 교차 상관관계
b) vs. M: 원본 파형과 로그-멜 규모 스펙트로그램으로부터 복원된 파형 사이의 교차 상관관계(SpecGAN 방식)
c) vs. M+IF: 원본 파형과 로그-멜 규모 스펙트로그램 및 멜 순시 주파수로부터 복원된 파형의 사이의 교차 상관관계 (GANSynth 방식)
d) vs. M+IF+HPSS: 원본 파형과 원본 파형을 HPSS를 통해 고조파 및 퍼커시브 파형으로 분해한 뒤 각각에 대하여 3)과 동일한 방식으로 복원한 파형 사이의 교차 상관관계 (본 논문에서 제안하는 방법)

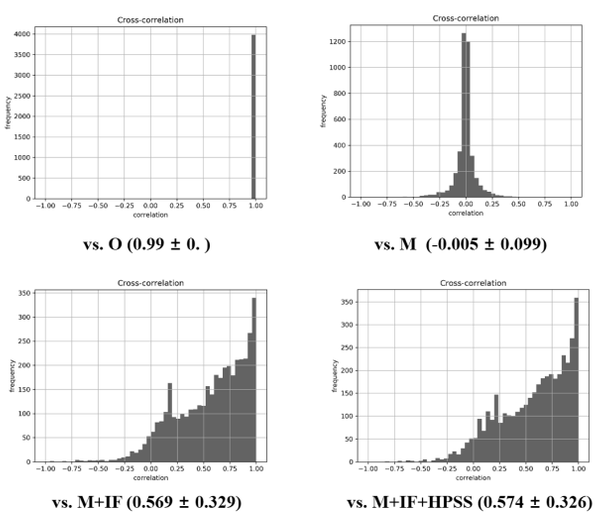
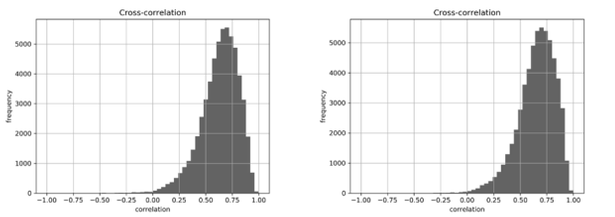


Fig. 4. Cross-correlations between Original Waveforms and Reconstructed Waveforms(NSynth Dataset)

SpecGAN에서 사용한 M은 -0.005±0.099로, GANSynth에서 제안한 M+IF는 0.569±0.329로, 본 논문에서 가설로 설정한 M+IF+HPSS 방식은 0.574±0.326으로 교차 상관관계가 측정되었다. M+IF+HPSS 방식이 평균은 소폭 증가 하였으나 통계적으로 유의미한 차이는 아니다. 따라서, NSynth Dataset과 같은 악기 소리에 대해서는 본 논문의 가설과 같이 정보의 손실을 줄이는데 도움을 주기 어렵다.

다음으로 Speech Commands Dataset에서 학습 데이터로 레이블링된 57,886개의 파형에 대한 상관관계를 측정한 결과는 Fig. 5와 같이 나타났다.



vs. M+IF (0.631 ± 0.18) **vs. M+IF+HPSS (0.653 ± 0.176)**

Fig. 5. Cross-correlations between Original Waveforms and Reconstructed Waveforms(Speech Commands Dataset)

M+IF는 0.631±0.18로, M+IF+HPSS는 0.653±0.176으로 측정되었다. 이 측정 결과는 통계적으로 유의미한 차이로 볼 수 있다. 따라서 음성과 같이 고조파 및 퍼커시브 성분이 고르게 나타나는 파형은 본 논문에서 제시한 가설과 같은 방식으로 파형을 복원할 경우 원본과 더 유사하다고 볼 수 있다.

4. 실험 및 성능평가

본 논문의 3.3절에서는 음성 파형에 대하여 본 논문의 가설 적용 유무에 따라 측정된 교차 상관관계가 통계적으로 유의미한 차이가 있음을 보였다. 본 장에서는 이 통계적 차이가 실제로 GAN이 생성하는 데이터의 충실도를 개선하는지 검증한다. 본 실험의 대조군은 GANSynth로 설정하였으며, 본 논문이 제안하는 GAN(실험군)은 4.1절에서 구조를 자세하게 다룬다. 4.2절에서는 생성된 데이터의 충실도와 다양성을 평가하기 위한 지표를 정의하고, 4.3절에서는 학습 환경과 학습 결과에 대하여 설명한다. 마지막으로 4.4절에서는 성능평가 결과에 대하여 논한다.

4.1 제안하는 GAN의 구조

GANSynth는 안정적으로 고해상도 스펙트로그램을 생성할 수 있음을 입증한 연구이다. 본 논문에서 제안하는 소리 생성 방법은 GANSynth의 구조를 따르지만, 3.3절의 교차 상관관계 측정 결과에 근거, M+IF+HPSS 방식을 기존 GANSynth 적용한 차이점이 있다. Fig. 6는 본 논문에서 제안하는 GAN에 대한 전반적인 구조를 담고 있으며, 구조에 대한 설명은 다음과 같다.

a) 생성기에 입력하는 잠재 벡터(latent vector)는 구형 가우시안(spherical Gaussian)으로부터 임의로 추출된 256차원 벡터와 원-핫 인코딩(One-hot encoding) 레이블을 벡터 결합(concatenate)하여 286차원 벡터로 구성한다.

b) 판별기에 입력하는 실제(real) 데이터는 본래 파형을 3.3절의 HPSS를 통해 고조파와 퍼커시브 파형으로 분해하고, 각 파형에 대하여 3.1절의 시간-주파수 표현을 사용하여 로그-멜 규모 스펙트로그램과 멜 순시 주파수로 변환한다. 이렇게 변환된 네 종류의 시간-주파수 표현은 수직으로 쌓아(stack) (128, 1024, 4) 쉐입의 텐서(Tensor)로 표현한다.

c) 생성기와 판별기 네트워크 모두 의 저해상도의 시간-주파수 표현을 시작으로 7단계의 성장을 거쳐 의 고해상도 시간-주파수 표현을 학습한다. 생성기의 업 샘플링 과정은 전치 컨볼루션을 통해, 판별기의 다운 샘플링은 평균 풀링(average pooling)을 통해 이루어진다. 생성기와 판별기 모두 활성 함수(activation function)로 리키 렐루(Leaky ReLU)를 사용한다. 그리고 생성기는 Equation (4)의 픽셀 정규화(pixel normalization)[13]를 추가적으로 적용한다. 해당 수식의 h, w, c는 각각 높이(주파수 축), 너비(시간 축), 채널의 차원을 지칭한다. x는 활성 함수를 통과한 특징 맵, C는 채널의 수를 의미한다.

$$x = x_{hwc} / \sqrt{\frac{1}{C}\sum_{c} x_{hwc}^{2} + \epsilon} \tag{4}$$

픽셀 정규화를 사용하는 것이 생성되는 데이터의 충실도를 높이는 직접적인 요인은 아니라고 알려져 있으나, 이미 학습된 네트워크에 새로운 네트워크를 쌓는 네트워크 성장 과정에서 발생하는 쇼크(shock) 현상을 완화시키는 것으로 알려져 있다.

d) 생성기의 출력 레이어에 사용한 활성 함수는 하이퍼볼릭 탄젠트(hyperbolic tangent)이며, 판별기의 출력 레이어는 소프트맥스(Softmax)를 활성 함수로 사용한다.

$$\begin{aligned} &\min_{G}\max_{D} V(D,G) \\ &= E_{x\sim p_{data}}[\log D(x|y)] + E_{z\sim p_{z}(z)}[\log(1-D(G(z|y)))] \end{aligned} \tag{5}$$

본 논문에서 제안하는 GAN은 Equation (5)와 같은 목적 함수(objective function)가 균형을 이루도록 학습이 진행된다. 본 목적 함수는 일반적인 GAN이 사용하는 목적 함수에 부가(auxiliary) 정보인 레이블 y가 조건부(conditional)로 추가되었다.

4.2 평가 지표(Evaluation Metrics)

GAN이 생성한 데이터가 얼마나 현실의 데이터와 닮았는지 공식화(formulation)하는 것은 매우 어려운 일이다. 물론 생성된 데이터를 사람이 직접 보거나 들어서 평가할 수 있겠으나, 주관적이며 현실적으로 시간과 비용이 많이 필요한 단점이 있다. 그래서 GAN 연구자들은 다양한 관점의 평가 지표을 제안하였고, 지표에 따라 장단점이 존재한다[18].

본 논문에서는 두 가지 측면에서 제안한 GAN을 평가한다. 먼저 생성된 데이터의 다양성이다. 다양성은 본 논문의 주된 평가 지표는 아니다. 그러나 다양성이 낮다는 것은 GAN이 학

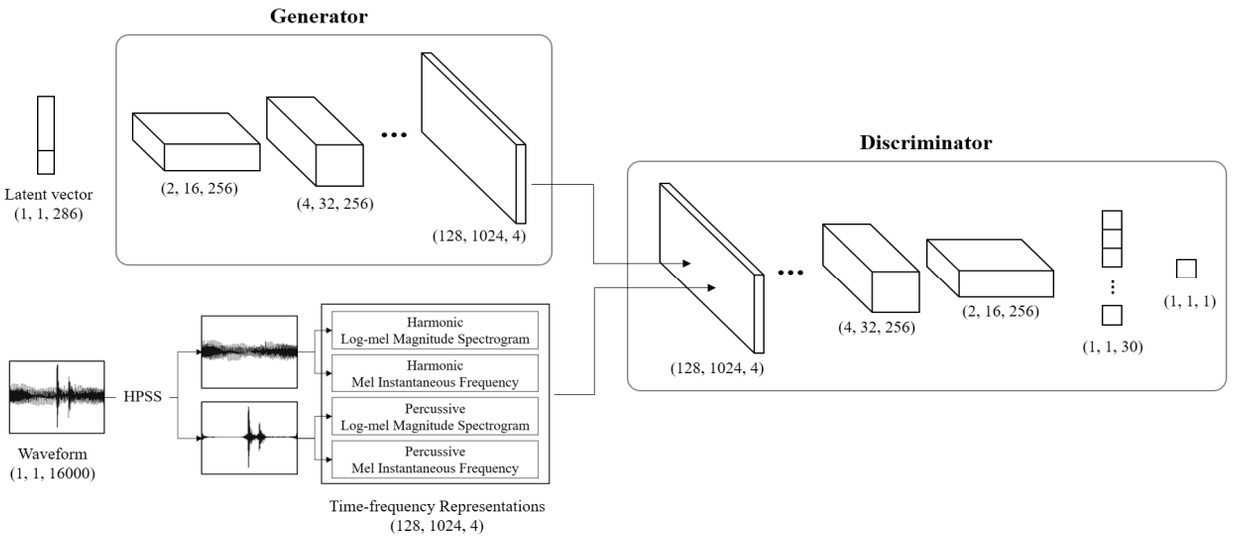


Fig. 6. An Illustration of our Proposed Method. The Generator Generates four Kinds of Fake Time-frequency Representations, which are Stacked in Sequence Vertically. The Discriminator Discriminates Real Representations from the Fake Ones. In Order to Learn High-resolution Representations, Both Generator and Discriminator Networks Grow as Training Progresses Progressively

습 과정에서 모드 붕괴(mode collapse)를 겪었음을 의미한다. 즉, 학습된 GAN이 특정한 소리 데이터만 생성하는 극도의 편향성이 나타난다. 이 모드 붕괴를 확인하기 위하여, 본 논문에서는 NDB (Number of Statistically-Different Bins)[19]를 사용한다. NDB는 학습 데이터에 대해 k-means를 이용해 k개의 군집을 이루는 모델을 학습 시키고, 테스트 데이터와 생성된 데이터를 근접한 군집에 배치(assign)하여 일정 임계값 이상 비율의 차이가 발생한 군집의 수를 센다. NDB가 작을수록 생성된 데이터가 학습 데이터와 비슷한 다양성을 가진다고 볼 수 있다. 본 논문은 로그-멜 규모 스펙트로그램 공간에서 k-means를 이용해 50개의 보로노이 셀(Voronoi cell)로 군집을 이루는 모델을 다양성 평가에 사용했다.

다음으로 생성된 데이터의 충실도 이다. 충실도는 본 논문의 주된 지표이다. GAN 연구에서 생성된 데이터의 충실도를 측정하는 대표적인 지표는 인셉션 점수(Inception Score, IS)[20]가 있다. 일반적으로 IS는 1,000개의 클래스를 가진 120만개의 이미지를 학습한 인셉션 모델[21]을 사용해 평가한다. 생성된 데이터를 인셉션 모델에 통과시켜 나온 출력의 확률 분포와 주변 분포(marginal distribution) 사이의 쿨백-라이블러 발산(Kullback-Leibler divergence)를 계산하여 평가한다. 그러나 IS는 실제 데이터의 분포를 사용하지 않는 단점이 있으며, 평가에 사용하는 인셉션 모델은 이미지 분류를 목적으로 학습된 모델이기 때문에, 본 연구와 같이 소리 데이터를 다루는 연구의 평가를 위한 모델로 사용하기에는 적절하지 않다. 그래서 본 논문에서는 실제 데이터와 생성된 데이터 사이의 프레쳇 인셉션 거리(Fréchet Inception Distance, FID)[22]를 충실도를 평가하기 위한 지표로 사용한다. FID는 사전 학습된 CNN 모델에 실제 데이터와 생성된 데이터를 통과시키고, 모델 중간의 특징 맵을 Equation (6)과 같은 수식을 사용해 거리를 계산한다.

$$FID(r,g) = \|\mu_r - \mu_g\|_2^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{\frac{1}{2}}) \qquad (6)$$

여기서 (μ_r, Σ_r)과 (μ_g, Σ_g)는 각각 실제 데이터의 특징 맵과 생성된 데이터의 특징 맵에 대한 평균과 공분산이며, Tr은 대각 성분의 합을 의미한다. 따라서 FID가 0에 가까울수록 충실도가 높다고 볼 수 있다. 본 연구에서는 FID 측정을 위해 Speech Commands Dataset에 대하여 3.1절의 시간-주파수 표현을 학습한 ResNet[23]을 FID 평가를 위한 모델로 사용한다. 이 모델은 테스트 데이터에 대하여 93.6%의 정확도(accuracy)를 보였다. 그리고 본 논문에서는 이 학습된 ResNet의 마지막 특징 맵인 512 차원의 벡터를 FID 측정에 사용했다.

4.3 학습 데이터 및 모델 학습

본 연구는 공개된 학습 데이터세트인 Speech Commands Dataset을 학습 데이터로 사용했다. 이 데이터세트는 크라우드 소싱 방식으로 다수의 익명 참여자가 30개의 짧은 단어(클래스)를 발음한 오디오 파일의 집합이며, 각 오디오 파일은 1초 분량이며 16kHz로 샘플링 되어 제공된다. 본 연구에서는 57,886개의 오디오 파일을 학습 데이터로 사용했으며, 6,835개의 오디오 파일은 테스트 데이터로 사용했다. 학습 데이터는 GAN, FID 측정을 위한 ResNet, NDB 측정을 위한 k-means 모델을 학습하는데 사용하였고, 테스트 데이터는 FID 측정을 위한 ResNet의 성능을 확인하고, FID와 NDB의 기준(criteria)을 마련하는데 사용하였다.

본 연구에서 제안하는 GAN과 대조군인 GANSynth 모두 16GB GPU 메모리의 NVIDIA P100 GPU 한 개를 사용하여 학습했다. 두 모델 모두 배치 크기를 4로 설정했을 때 GPU 메모리의 95% 가량을 필요로 했다. 총 학습 시간은

300 에폭(epoch)을 돌파하는데 약 360시간(15일)이 소요되었다. 학습 과정에서 계산된 두 모델의 손실(loss) 값은 Fig. 7과 같이 나타났다. 0-40 에폭에서 발생한 손실 값의 큰 변화는 네트워크가 성장하는 과정에서 발생한 자연스러운 현상이며, 이후에 보여 지는 손실 값과 같이 진동(oscillation) 문제는 발생하지 않았다.

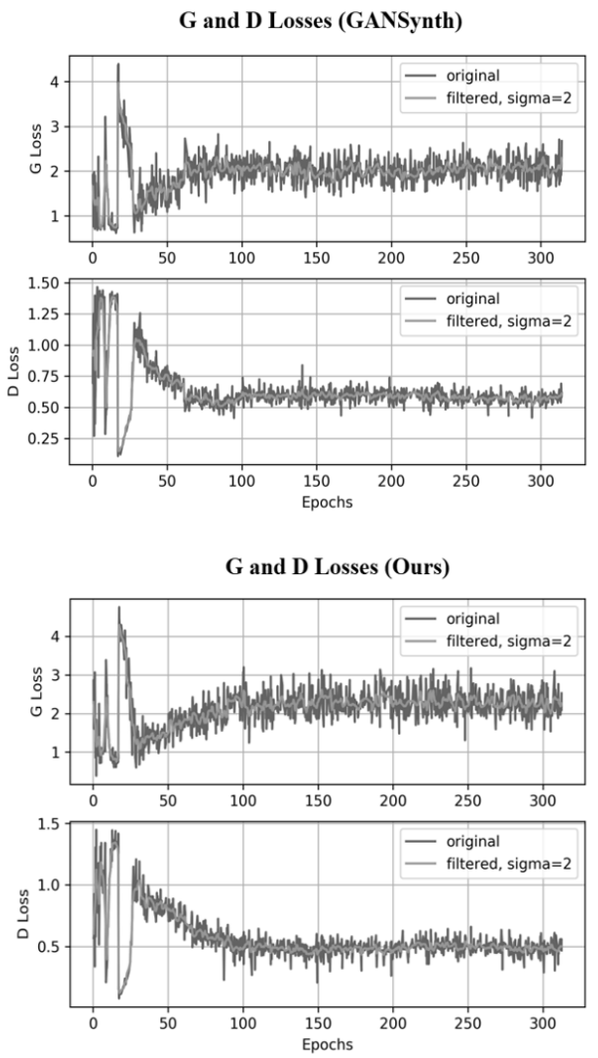


Fig. 7. Generator and Discriminator Losses. (Top) GANSynth. (Bottom) Ours

4.4 성능 평가

Table 1은 본 논문에서 제안하는 GAN을 4.2절에서 정의한 지표로 평가한 결과이다. FID는 임의로 추출한 3,000개의 샘플에 대하여 측정하고, 독립적으로 20회 측정한 FID의 평균과 표준편차를 산출하였다. NDB는 임의로 추출한 1,000개의 샘플에 대하여 측정하고, 독립적으로 10회 측정한 NDB의 평균과 표준편차를 산출하였다. 그리고 Table 1의 Real은 학습에 사용하지 않은(unseen) 테스트 데이터이며, 본 성능 평가의 기준점이다.

Table 1. Evaluation Metrics of our Method and the Baseline

Examples	Epochs	NDB	FID
Real	-	4.11±1.62	0.275±0.017
GANSynth	124	29.1±2.119	15.822±0.013
	172	22.0±1.897	14.869±0.131
	207	27.3±3.198	14.928±0.142
	241	17.9±2.587	13.68±0.107
	276	21.7±2.002	**11.973±0.141**
	310	**14.11±1.27**	12.565±0.146
Ours	124	28.0±2.757	14.924±0.174
	172	28.4±2.2	13.094±0.136
	207	31.2±1.833	11.04±0.093
	241	27.8±1.72	**10.504±0.112**
	276	27.7±2.9	12.32±0.119
	310	**20.44±1.59**	11.749±0.138

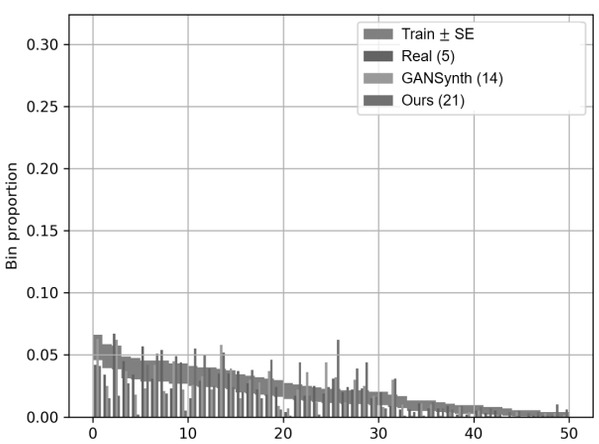


Fig. 8. Bin Proportions of the Generated Waveforms Assigned to the 50 Clusters Trained by k-means in Log-spectrogram Space

먼저 다양성 측면에서 성능 평가 결과를 살펴보면, 본 논문이 제안하는 GAN은 약 20개, GANSynth는 약 14개의 군집에서 학습 데이터보다 많은(또는 적은) 비율의 데이터를 생성했다. Real의 NDB가 약 5로 측정되었음을 감안하면 다양성이 낮다고 볼 수 있으나, Fig. 8의 대표 히스토그램과 같이 두 모델 모두 특정 군집에 과도하게 편향된 결과는 아니다. 즉, 두 모델 모두 모드 붕괴가 발생했다고 보긴 어렵다.

다음으로 충실도 측면에서 평가 결과를 살펴보면, 본 논문이 제안하는 GAN은 10.504±0.112(241 epochs)의 FID를 보였으며, GANSynth는 11.973±0.141(276 epochs)의 FID를 보였다. 두 모델 모두 학습이 진행됨에 따라 FID가 낮아지지만, 일관되게 유의미한 FID 차이가 발생했다. 따라서 본 논문에서 설정한 가설은 GAN이 생성한 음성 데이터의 충실도를 개선하는데 영향을 주었다고 해석할 수 있다. 본 논문

이 제안하는 GAN은 GANSynth 보다 높은 충실도(낮은 FID)를 보였으나, Real의 FID인 0.257±0.007과 비교하면 아직 개선의 여지가 많다(실제로 두 모델이 생성한 음성 데이터는 사람의 청각으로 평가했을 실제과 가상을 비교적 쉽게 구분할 수 있다). 따라서, GAN을 통해 사람의 청각으로도 실제와 가상을 구분하기 어려운 수준의 음성을 생성하기 위해서는 추가적인 연구가 필요함을 시사한다.

5. 결론 및 향후 연구

본 논문에서는 소리 데이터의 범위를 음성으로 한정하여 GAN이 생성한 음성의 충실도를 개선하는 방법을 제안하였다. 개선된 방법은 음성에 고조파와 퍼커시브 성분이 고르게 나타난다는 특성에 착안하여, HPSS를 이용해 주어진 파형을 고조파와 퍼커시브 파형으로 분해하고, 각 파형을 시간-주파수 표현으로 변환하여 GAN의 학습 데이터로 사용하는 방식이다. 본 논문에서 제안하는 방법이 적용된 GAN은 공개된 음성 데이터세트로 학습 시켰으며, NDB를 통해 다양성을 평가하고, FID를 사용하여 충실도를 평가했다. 성능평가 결과에 따르면, 본 논문에서 제안하는 GAN은 기존의 최신 소리 생성 GAN 보다 생성된 음성의 충실도가 높게 나타났다.

본 논문에서 제안하는 방법은 최신 연구보다 높은 충실도(낮은 FID)를 달성하였다. 따라서 데이터 증강 기법으로서, 특별한 데이터 수집 없이 기존의 학습 데이터를 활용하여 음성 인식과 같은 판별 모델의 성능을 개선하는데 활용될 수 있다. 특히, 주어진 학습 데이터의 불균형 문제를 개선하거나, 의료 분야와 같이 매우 희소한 데이터를 사용하여 판별 모델을 개발하기 위한 목적으로 활용될 수 있을 것이다.

향후 연구에서는 복소 영역에서 스펙트로그램을 학습할 수 있는 새로운 구조의 CNN을 설계하거나, 시간-주파수 표현을 사용하지 않고 직접적으로 파형을 학습 하고 생성하는 순환 신경망(Recurrent Neural Networks) 구조의 GAN을 설계하는 방향으로 본 논문을 확장할 예정이다.

References

[1] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp.1097-1105. 2012.

[2] ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [Internet], http://www.image-net.org/challenges/LSVRC/

[3] Detection and Classification of Acoustic Scenes and Events (DCASE) [Internet], http://dcase.community/

[4] TensorFlow Speech Recognition Challenge [Internet], https://www.kaggle.com/c/tensorflow-speech-recognition-challenge

[5] I. Goodfellow, et al,. "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, pp.2672-2680, 2014.

[6] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[7] Y. Wu, J. Donahue, D. Balduzzi, K. Simonyan, and T. Lillicrap, "Logan: Latent optimisation for generative adversarial networks," *arXiv preprint arXiv:1912.00953*, 2019.

[8] D. Nie, et al., "Medical image synthesis with context-aware generative adversarial networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp.417-425, 2017.

[9] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[10] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," *arXiv preprint arXiv:1802.04208*, 2018.

[11] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, Vol.1, No.10, pp.e3, 2016.

[12] J. Engel, K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "Gansynth: Adversarial neural audio synthesis," *arXiv preprint arXiv:1902.08710*, 2019.

[13] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[14] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, "Neural audio synthesis of musical notes with WaveNet autoencoders," in *International Conference on Machine Learning*, pp.1068-1077, 2017.

[15] M. Tayyab, I. Ahmad, N. Sun, J. Zhou, and X. Dong, "Application of integrated artificial neural networks based on decomposition methods to predict streamflow at Upper Indus Basin, Pakistan," *Atmosphere*, Vol.9, No.12, pp.494, 2018.

[16] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proceedings of the International Conference on Digital Audio Effects (DAFx-10)*, pp.217-220, 2010.

[17] P. Warden, "Speech commands: A public dataset for single-word speech recognition", *Dataset available from http://download.tensorflow.org/data/speech_commands_v0.01.tar.gz*, 2017.

[18] A. Borji, "Pros and cons of gan evaluation measures," *Computer Vision and Image Understanding*, Vol.179, pp.41-65, 2019.

[19] E. Richardson, and Y. Weiss, "On gans and gmms," in *Advances in Neural Information Processing Systems*, pp.5847-5858, 2018.
[20] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *arXiv preprint arXiv:1606.03498*, 2016.
[21] C. Szegedy, et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1-9, 2015.
[22] 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," in *Advances in Neural Information Processing Systems*, pp.6626-6637, 2017.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770-778, 2016.

백 문 기
https://orcid.org/0000-0001-7695-970X
e-mail : zmzment@gmail.com
2013년 충남대학교 컴퓨터공학과(학사)
2013년 ~ 현 재 충남대학교 컴퓨터융합학부 석·박사통합과정
관심분야 : 딥러닝, 생성모델, 음성인식, 신호처리

윤 승 원
https://orcid.org/0000-0003-1366-9620
e-mail : yoonenoch11@gmail.com
2018년 충남대학교 컴퓨터공학과(학사)
2018년 ~ 현 재 충남대학교 컴퓨터융합학부 석·박사통합과정
관심분야 : 딥러닝, 생성모델

이 상 백
https://orcid.org/0000-0002-4440-1792
e-mail : roy881020@gmail.com
2014년 충남대학교 컴퓨터공학과(학사)
2014년 ~ 현 재 충남대학교 컴퓨터융합학부 석·박사통합과정
관심분야 : 딥러닝, Human-object interaction

이 규 철
https://orcid.org/0000-0003-0857-807X
e-mail : kclee@cnu.ac.kr
1984년 서울대학교 컴퓨터공학과(학사)
1986년 서울대학교 컴퓨터공학과(석사)
1990년 서울대학교 컴퓨터공학과(박사)
1990년 ~ 현 재 충남대학교 컴퓨터융합학부 교수
1994년 미국 IBM Almaden Research Center 초빙연구원
1994년 미국 Syracuse University, CASE Center 초빙교수
관심분야 : 데이터베이스, 딥러닝, 빅데이터, 융합기술