공학석사 학위논문

DeepASMR: 딥러닝 기반의 ASMR 플랫폼

2020 년 12 월

서울대학교

전기정보공학부

문 지 영

DeepASMR: 딥러닝 기반의 ASMR 플랫폼

지도교수 고 형 석

이 논문을 공학석사 학위논문으로 제출함
2020년 12월

서울대학교
전기정보공학부
문 지 영

문지영의 석사 학위논문을 인준함
2020 년 12 월

위 원 장 최 진 영 (인)

부위원장 고 형 석 (인)

위 원 김 영 민 (인)

초 록

최근 ASMR(Autonomous Sensory Meridian Response)이라고 불리는 심리적 안정감을 제공하는 “특별한” 소리에 관심이 높아지고 있으며 관련 비즈니스도 활발히 시작되고 있다. 그러나 ASMR 음원 개발은 많은 시간과 노력이 필요하여 생산성 문제가 있다. 본 연구는 딥러닝을 이용하여 기존 ASMR 음원들을 모으고 분류하며 사용자의 선호도를 바탕으로 새로운 ASMR 음원을 생성할 수 있는 플랫폼인 *DeepASMR*을 제안한다. DeepASMR은 ASMR 음원 분류 및 인식을 위해 기존의 음악 인식이나 소음 인식을 위한 DNN보다 개선된 DNN 모델들을 구축하여 분류의 정확도를 95% 이상까지 높였다. 이를 통해 인터넷에 있는 ASMR 음원들을 분류하여 데이터베이스를 만들고, 사용자에게 자극(trigger)이 되는 ASMR을 쉽게 찾을 수 있도록 한다. 또한 DNN을 기반으로 기존 ASMR 음원들을 변형하거나 합성하는 방식으로 새로운 ASMR 음원을 생산한다. 이를 위해 VAE(Variational Autoencoder) 및 GAN(Generative Adversarial Network) 방식을 이용하여 ASMR 음원 생성 DNN 모델을 구축하였다. 이를 통해 생성된 ASMR 음원들을 우리의 분류 DNN 모델에 입력하여 그 정확성을 검증한 결과, 70% 이상의 정확도를 보여 제안하는 DNN 모델이 양질의 ASMR 음원들을 생성하였음을 시사한다.

주요어 : ASMR, 딥러닝, 분류, 생성, VAE, GAN
학 번 : 2019-24791

목 차

표 목 차

그 림 목 차

제 1 절 서 론

최근 유튜브에서는 사탕먹는 소리나 속삭이는 소리 등 매우 “특별한 소리”를 찾는 구독자들이 늘고 있다. 이러한 소리는 *ASMR*(Autonomous Sensory Meridian Response, 자율 감각 쾌락 반응)을 주는데, 이는 청각 등 다양한 자극(trigger)을 통해 사람이 느끼게 되는 심리적인 안정감을 의미한다[1]. ASMR은 인간의 뇌에 엔도르핀과 옥시토닌의 분비를 촉진시켜 집중력 향상과 수면 유도 등의 효과를 보이며, 이를 규명하기 위해 심리학과 정신물리학 등에서 활발히 연구가 진행되고 있다[1]. ASMR에 대한 네티즌의 관심도 매우 높아(특히 한국) 지난 10년간 약 1억 2천만 회 이상의 구글 비디오 검색이 이루어졌으며 최근에 상용화된 ASMR 비디오와 광고 및 앱 들이 활발히 출시되고 있다[9,14].

그러나 이러한 관심과 효용성에도 불구하고 ASMR 음원의 개발과 활용은 매우 느리고 비체계적이다. ASMR 음원 제작을 위해서는 조용한 스튜디오에서 고가의 장비와 재료 및 많은 시간과 노력이 필요하다. 또한, 개인마다 자극을 느끼는 ASMR이 매우 다르므로 보편적인 ASMR 개발이 쉽지 않아 ASMR 음원의 대량 생산이 요구된다. 기존 심리학에서 진행하는 ASMR 연구도 개인 표본을 모집할 때 대부분의 ASMR에 반응하는 표본들이 선택되어 일반화하는 데 어려움을 겪는다[1]. 특히 자극을 느끼는 ASMR 음원을 발견해도 이를 계속 들으면 쉽게 싫증이 나므로 유사한 ASMR 음원들의 생성도 필요할 수 있다. 따라서 ASMR 음원 개발의 생산성을 획기적으로 높이는 새로운 플랫폼이 요구된다.

본 연구에서는 딥러닝을 기반으로 한 ASMR 음원의 분류 및 생성 플랫폼인 *DeepASMR*을 제안한다. DeepASMR에서는 (1) DNN을 이용한 분류기(classifier)를 통해 인터넷에 존재하는 ASMR 음원들을 모으고 분류하여 ASMR 음원 데이터베이스를 구축하여 사용자들에게 자극을 느끼는 음원들의 선택을 도와준다. (2) 선택된 ASMR 음원들에 대해 DNN

을 이용하여 다양하게 변형되고 합성된 유사한 음원들을 생성한다.

본 연구에서 (1)의 해결을 위해 사람이 다양한 ASMR 소리를 어떻게 구분해서 인지하는지를 DNN을 통해서 구현한다. 구체적으로 소리를 주파수를 표현하는 MelSpectrogram으로 변형한 후 그 특성을 임베딩 공간에서 표현하고 이를 분류한다. ASMR 음원은 기존의 생활 소음이나 음악과는 유사한 점도 있지만, 매우 다른 특성들도 존재하고 있다. 따라서 기존의 음악을 인식하는 DNN[4,7] 혹은 소음을 인식하는 DNN[2]과는 다른 DNN 모델들을 제안한다. 이를 통해 ASMR 음원 인식의 정확도를 95%까지 올릴 수 있음을 실험으로 보였다.

또한, 본 연구에서는 (2)의 해결을 위해 (1)에서 구축한 DNN들을 기반으로 ASMR 음원 생성모델을 구축하였다. (1)의 DNN을 인코더와 디코더로 사용하는 단순 오토인코더(Autoencoder) 모델과 이를 확장한 VAE(Variational Autoencoder)[10] 모델을 사용하여 변형된 음원의 MelSpectrogram을 생성한다. 오토인코더 모델은 기존 음원을 거의 변형하지 못하는 데 비해서 VAE 모델을 좀 더 변형된 음원을 생성하거나 두 가지 ASMR 음원을 합성하는 데 유리하다. 하지만 기존 VAE의 문제점 중 하나로 알려진 Posterior Collapse 문제[17]가 ASMR 음원에 대해서도 발생함을 확인할 수 있었다. 그래서 또 다른 생성 DNN 모델인 GAN(Generative Adversarial Network)[11] 방식을 채택한 생성모델도 구축하여 변형과 합성을 통해 새로운 음원을 생성하였다. 본 논문에서는 이렇게 생성된 ASMR 음원들을 (1)에서 구축한 분류 DNN의 입력으로 넣어 음원들의 정확성을 평가하였고 또한 음원의 결과를 데모와 함께 보고한다.

본 저자가 아는 한 DeepASMR은 DNN 기반의 ASMR 음원 분류 및 생성에 대한 첫 번째 연구 시도이며, 따라서 다양한 도전적인 문제들을 제시하고 그에 대한 해법들을 제안했다는 점이 본 연구의 공헌이다. 현

재 ASMR에 대한 관심이 높아지고 관련 온라인 비즈니스가 커지는 상황에서 DeepASMR이 이 분야의 발전에 기여할 것으로 기대한다.

제 2 절 배 경 지 식

2.1 딥러닝을 이용한 분류

기존 학계에서는 ASMR의 효과가 입증되지 않았기 때문에 ASMR 소리 분석에 대한 연구가 거의 이루어지지 않았다. 본 연구는 ASMR 소리와 가장 유사한 소리를 연구하는 생활 소음 분야(Environmental Sound)의 연구를 참고하여 진행하였다[2].

ASMR 소리와 생활 소음은 음악적으로 반복되는 구조나 박자가 없고 사람이 특정 소리를 인지하는 데 소요되는 시간이 짧다는 유사한 특징을 지닌다. 소음 분야에서는 이러한 특징을 잘 학습하고 분류할 수 있는 1D CNN 모델이 많이 사용된다[2]. 이는 음악 혹은 음성인식에서 많이 사용하는 RNN 계열의 딥러닝과 달리 소음에는 현재와 먼 과거 데이터 사이의 연관성이 적기 때문이며 이는 ASMR 소리도 유사할 것이다. 그러나 생활 소음과 ASMR의 차이점도 존재한다. 구체적으로 ASMR은 일상 소음 달리 잡음이 없고 매우 조용한 환경에서 녹음이 이루어진다. 볼륨이 작고 섬세하며 생활 소음 카테고리와 거의 겹치지 않고 한 가지 물체의 소리라도 생활 소음에 비해서 상대적으로 다양한 음색(timbre)을 갖추고 있다는 점에서 상이하다. 따라서 기존의 생활 소음에 사용된 분류기만으로는 ASMR 소리를 적절히 구분하기 어려울 수 있다.

2.2 딥러닝을 이용한 생성

DNN을 이용한 생성 모델(Generative Model)은 기본적으로 분류모델(Discriminative Model)보다 개발하기 어렵다. 이는 더 많은 정보를 학습해야 하기 때문이다. 이미지의 생성에 비해 ASMR과 같은 소리의 생성 결과는 소음이 많이 포함될 수 있어서 그 정확성을 판단하기 더 어렵다. 특히 개인의 경험과 소리에 대한 추측 등이 상대적으로 결과 판단에 더 많은 영향을 주기 때문에, 소리의 생성모델 구축은 더 도전적이다.

기존 이미지 생성 모델로는 단순 오토인코더(Autoencoder) 방식, VAE(Variational Autoencoder) 방식[10], 그리고 GAN(Generative Adversarial Network)[11] 방식이 많이 사용되고 있다.

오토인코더(AE)는 입력의 차원을 줄이는 인코더와 차원을 다시 늘리는 디코더로 구성되며 학습 시에 입력과 출력에 동일한 데이터가 사용된다. 이는 입력 데이터의 차원을 줄이는데 혹은 입력의 패턴을 학습하는데 유용한 구조이다. 인코딩이 진행됨에 따라 정보의 손실이 발생하므로 유사한 소리의 생성을 기대할 수 있다.

VAE는 오토인코더와 구조적으로 유사하다. 차이점으로는 인코더 이후에 샘플링 레이어를 추가함으로써 인코딩된 결과를 샘플링한 결과를 디코더에 입력한다. VAE는 AE와 다르게 입력의 확률분포를 추정하는 것을 목표로 한다. 즉 입력만 사용하는 것이 아니라 Variational Inference 방법을 사용해서 입력의 확률분포를 추정하는 역할을 샘플링 레이어가 수행한다. 그림 1은 VAE의 개념도를 보여준다.

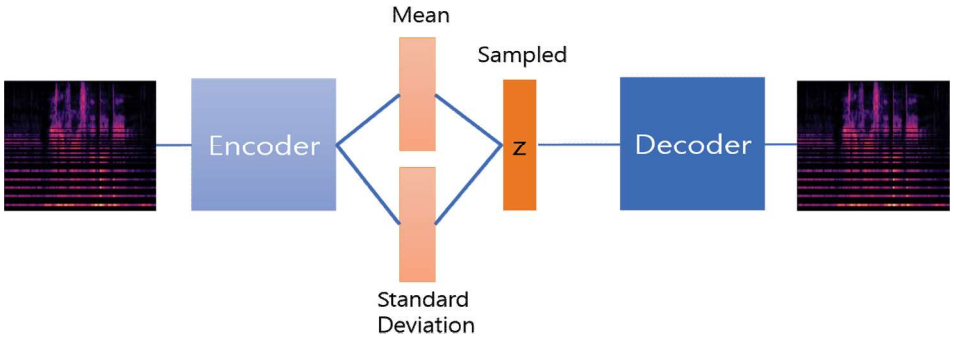


그림 1. VAE 모델의 개념도

GAN(Generative Adversarial Networks)은 적대적 생성 신경망으로 VAE처럼 입력의 확률분포를 추정하는 것을 목표로 한다. 이는 노이즈로부터 이미지를 생성하는 Generator와 이미지의 사실 여부를 판단하는 Discriminator로 구성된다. Generator는 Discriminator가 구분하기 어렵게 진실된 이미지와 유사한 이미지를 생성하는 것을 목표로 하고, Discriminator는 이에 대응하여 진실된 이미지와 Generator에 의해 생성된 이미지를 구분하는 것을 목표로 한다. 이 둘은 서로 경쟁을 통해 학습을 진행하면서 서로의 성능이 향상시키며 결과적으로는 간접적으로 입력의 확률분포를 학습하게 된다. 그림 2(a)는 GAN의 개념도를 보인다.

ASMR 소리 생성을 위해서는 레이블을 함께 모델의 입력으로 사용하는 cGAN(Conditional GAN)[15]이 더 적절하다. 일반 이미지에 비해 MelSpectrogram은 레이블 없이 결과를 눈으로 바로 구분하기 어렵기 때문에 레이블 정보를 함께 사용하는 것이 유용하다. 또한 서로 다른 두 소리를 결합하는 과정에서도 임베딩된 두 레이블 값의 평균을 입력값으로 사용할 수 있어 편리하다. 모델에 레이블로 Condition을 주는 방식으로 원하는 레이블의 결과를 얻을 수 있다. 그림 2(b)는 cGan의 개념도를 보여준다. 이러한 cGAN에서 이미지 대신 ASMR 음원의 MelSpectogram을 사용하면 ASMR 음원을 생성할 수 있을 것이다.

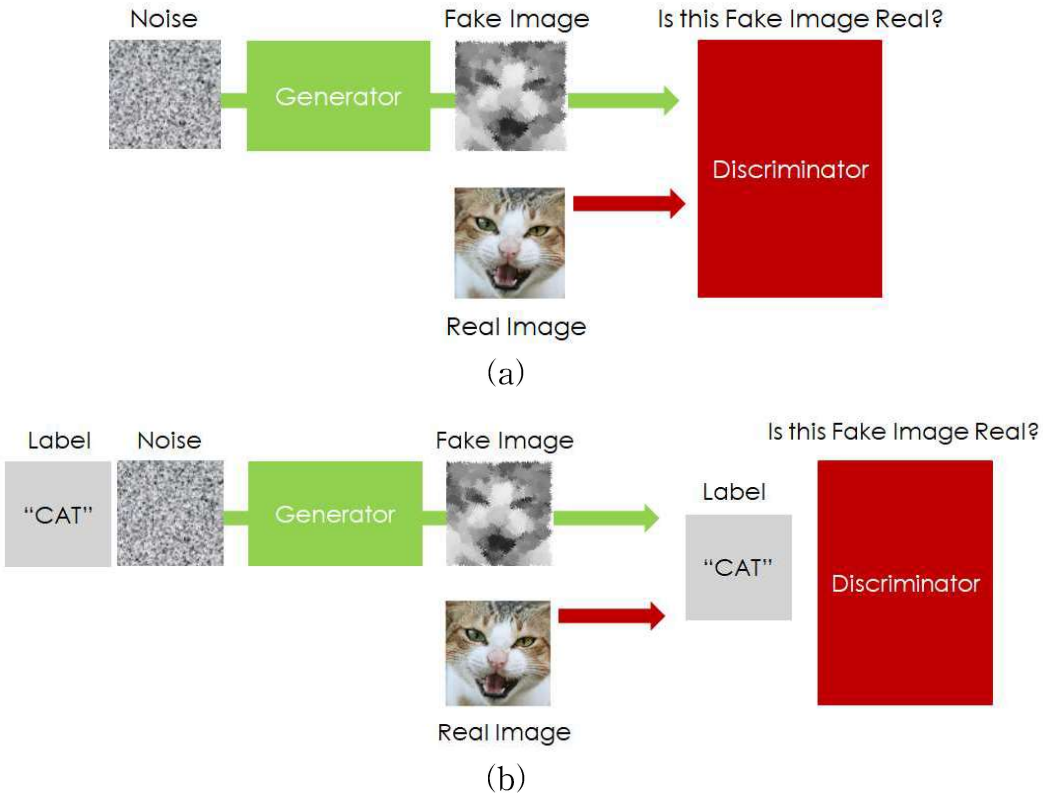


그림 2. GAN 모델(위)과 cGAN모델(아래)의 개념도

GAN와 VAE 두 모델은 결과적으로 데이터의 확률분포를 학습한다는 공통점이 있다. VAE는 오토인코더 구조로 인코딩된 결과를 샘플링하여 디코딩한다. GAN은 생성자가 랜덤의 값을 통해 바로 결과를 만들어내며 구분자가 원본과 생성된 결과를 구분하기 어렵게 만든다.

GAN 기반 DNN 모델과 VAE는 다음과 같은 차이점이 있다. DNN 네트워크 면에서 VAE는 인코더를 사용한 다음 디코더를 적용하는 데 비해 GAN의 경우 디코더에 해당하는 Generator를 사용한 다음에 인코더에 해당하는 Discriminator를 적용한다고 볼 수 있다. 디코더 입력의 경우에도 VAE는 입력으로부터 랜덤하게 샘플된 데이터를 사용하고 GAN의 경우는 랜덤하게 샘플된 노이즈를 사용한다. 디코더의 출력의 경우 VAE는 입력 데이터에 충실하지만 희미(blurry)하고 GAN의 경우는 러

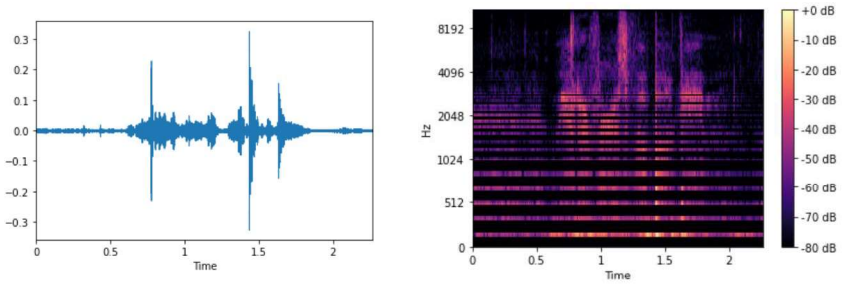
닝에 사용된 데이터들과 유사하고 더 선명하다.

제 3 절 딥러닝을 이용한 ASMR 분류

본 절에서는 우리가 ASMR 분류를 위해 어떻게 ASMR 음원을 처리했는지와 ASMR 인식을 위한 어떤 DNN 모델 설계하고 구축했는지를 설명한다.

3.1 ASMR 음원 처리

인터넷에 있는 ASMR들을 모아서 데이터베이스를 구축하기 위해 세 가지 과정을 거쳤다. 우선 유튜브 인기 ASMR 영상 중 9가지 장르의(부서지는 모래, 구겨지는 플라스틱 봉지, 손톱으로 사물을 태핑, 속삭이는 남자, 속삭이는 여자, 말하는 남자, 쭉 늘리는 슬라임, 마이크에 붓질, 먹방) 소리 레이블을 정한 후 30시간 분량의 유튜브 음성 파일을 모아 오디오를 추출한다. 그림 3(a)는 오디오 파형의 예이다.



(a) (b)

그림 3. 오디오와 MelSpectrogram의 예

이러한 오디오를 그대로 DNN의 입력으로 사용할 경우 Nyquist Limit에 의해 1초에 44100개의 값(44.1kHz)으로 샘플링되어 사용될 수 있다.

샘플링된 수 만개의 연속적인 데이터는 시간에 따른 소리의 세기 변화만을 나타낸다. 이 자체로는 소리의 높낮이에 해당되는 주파수에 대한 정보는 제공하지 않는다. 따라서 오디오를 기반으로 DNN을 구성하면 소리의 인식률이 높지 않다고 알려져 있으며 5 절에 있는 우리의 실험에서 보듯이 ASMR에 대한 인식률도 역시 높지 않았다.

소리의 특성을 제대로 파악하기 위해서는 시간, 소리의 세기 및 높낮이에 대한 정보가 모두 필요하다. 따라서 이 세 가지를 모두 표현할 수 있는 Spectrogram으로 오디오를 변환하였다. 더 정확히는 인간이 낮은 주파수 대역의 차이를 높은 주파수 대역보다 민감하게 반응하는 특성을 반영하기 위해 주파수에 MelScale을 사용하는 MelSpectrogram으로 변환하였다. 실제 딥러닝 기반의 음성인식에서도 MelSpectrogram이 많이 사용되고 있다. 그림 3의 (b)는 MelSpectrogram의 예를 보여준다. MelSpectrogram을 사용함으로써 기존 이미지에서 사용하는 CNN이나 Generative 모델들을 사용할 수 있고 원래의 오디오로 변환하면 다시 ASMR을 얻을 수 있다. 우리는 MelSpectrogram을 약 1초씩 나누어 분류 DNN 모델의 입력으로 사용하였다.

3-2. 분류 DNN 모델

본 절에서는 기존에 음악과 소음을 인식하는 데 사용했던 다양한 DNN 모델들을 ASMR 인식을 위한 대안으로 제시하고, 또한 우리가 제안하는 ASMR에 특화된 DNN 모델의 구조를 설명한다.

ⓐ 오디오를 기반으로 하는 DNN은 하나의 dense layer를 기반으로 하는 경우가 종종 사용된다(**single dense layer on audio**) [4,7].

ⓑ MelSpectrogram을 기반으로 하는 음악 인식 DNN은 2D CNN이 사용되는 경우가 많다(**2D CNN on MelSpectrogram**) [4,8].

ⓒ 2D CNN 하나를 사용하는 ⓑ 대신에 2D CNN 네 개를 사용하는 DNN을 구축하였다. 네 개를 사용한 이유는 ⓔ 에서의 이유와 같다. (**4**

Layers of 2D CNN on MelSpectrogram).
ⓓ 생활 소음을 인식하는 DNN의 경우 1D CNN이 효과적이라는 결과가 있다(**1D CNN on MelSpectrogram**) [2].
ⓔ 특히 네 개 layer의 1D CNN 모델을 사용하면 생활 소음을 더 효과적으로 인식한다는 관찰도 있다[2]. 우리는 이를 근거로 ASMR를 인식하기 위해 그림 4와 같이 네 개의 1D CNN 모델을 기반으로 MaxPool, Dropout, global averare pool(GAP), 및 dense layer를 적절히 추가하여 DNN을 구축하였다(**4 Layers of 1D CNN on MelSpectrogram**).
ⓕ 이 구조에 세부조절을 위해 activation 함수로 일반 ReLU 대신 ELU와 Tanh를 사용하고, RMSprop을 optimizer로 사용하며 다양한 세부 조절과 엔지니어링을 통해 최종적인 DNN을 구축하였다(**4 Layers of 1D CNN after fine tuning on MelSpectrogram**).

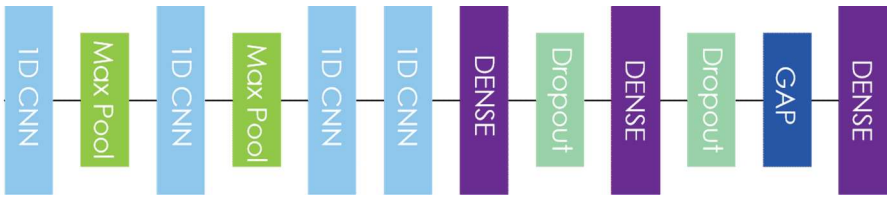


그림 4. 제안하는 ASMR 인식(분류) DNN 모델

이렇게 구축된 DNN들을 이용하여 3-1절의 ASMR 음원들의 인식 정확도를 평가한다. 그리고 4절에서는 이들을 기반으로 ASMR 음원 생성모델을 구축한다.

제 4 절 딥러닝을 이용한 ASMR 생성

본 장에서는 우리가 구축한 오토인코더 방식, VAE 방식, GAN 방식의 ASMR 음원 생성모델의 설계와 구현을 설명한다. ASMR 음원 생성 방법으로는 기존 ASMR 음원을 변형하는 방식과 두 가지의 ASMR 음원

을 합성하는 방식을 사용한다.

4-1. 오토인코더(AE) 기반의 생성모델

오토인코더(AE) 방식은 단순히 3절 ⓕ의 분류 DNN을 기반으로 인코더와 디코더를 사용하여 학습하는 생성모델이다. 그림 5는 우리가 적용한 오토인코더 모델을 보여준다. 이 모델은 인코딩에 따라 정보 소실이 발생하여 변형된 소리를 생성할 것으로 기대했으나 5절의 실험 결과에서 보듯이 매우 유사한 ASMR 음원을 생성하여 변형의 효과가 별로 나타나지 않아서 생성모델로 사용하기에는 제한적이었다.

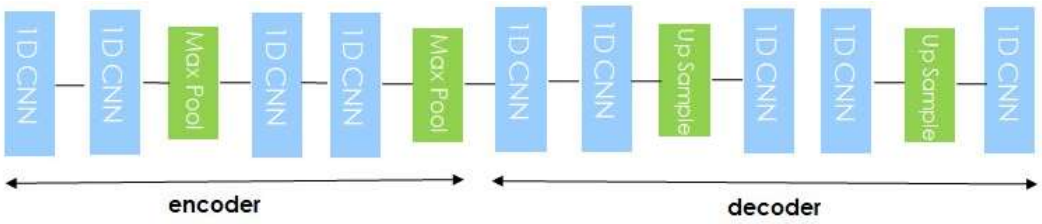


그림 5. 오토인코더(AE) 기반의 encoder/decoder ASMR 생성모델

4-2. VAE 기반 DNN 모델

오토인코더에 샘플링 레이어를 추가한 Variational Autoencoder(VAE)를 구축하면 소리의 확률 분포를 학습하기 때문에 좀 더 다양한 소리를 생성할 수 있다. 3절 ⓕ의 분류 DNN을 기반으로 그림 6과 같이 VAE를 구성하였다. 역시 동일한 ASMR 음원의 MelSpectrogram을 모델의 입력과 출력으로 가지고 학습하여 변형된 ASMR 음원을 생성하도록 하였다. 우리는 Latent Dimension을 256으로 지정하였다.

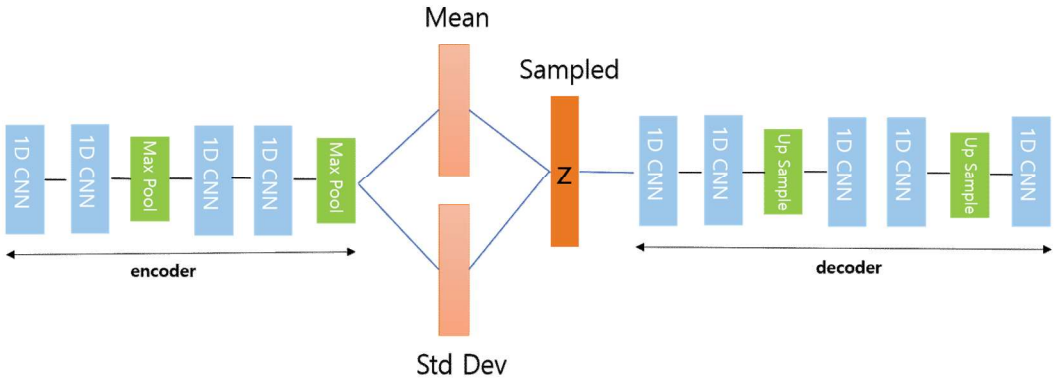


그림 6. VAE 기반의 encoder/decoder ASMR 생성모델

디코더 부분에 레이어를 추가할수록 결과가 향상되었다. 기존에 효과적이라고 알려진 네 개의 레이어를 사용했을 때[3,6]와 비교했을 때, 다섯 개의 레이어를 사용한 모델의 결과가 향상되는 것이 보였다. 하지만 다섯 개를 초과해서 레이어를 사용할 경우 다시 학습이 잘 이루어지지 않았다.

위의 VAE를 기반으로 두 가지 ASMR 음원을 합성하는 모델을 그림 7과 같이 구축하였다. 두 가지 다른 소리에 대해 각각 인코딩 후 평균을 낸 음원을 디코딩하면 두 소리의 합성된 음원도 얻을 수 있다.

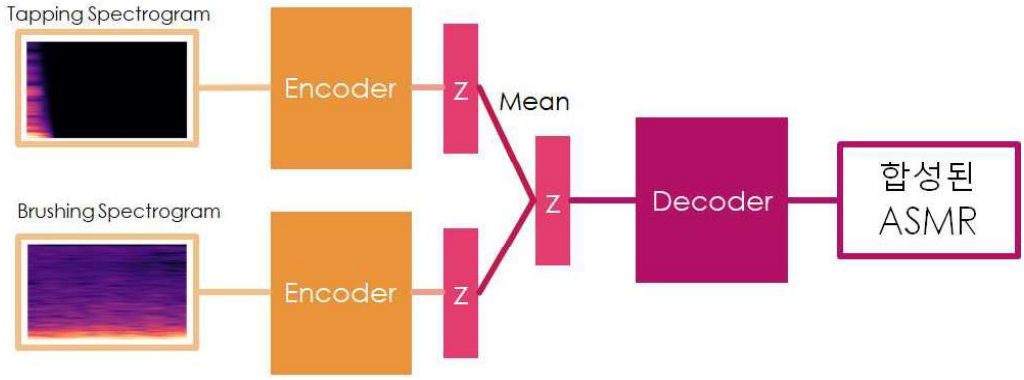


그림 7. VAE 기반의 ASMR 음원의 합성 생성모델

4-3. GAN 기반 DNN 모델

우리는 2절의 그림 2(b)와 같이 cGAN 기반의 DNN 모델을 구축하였다. 그림 8은 구축된 cGAN 모델의 Generator와 Discriminator의 DNN을 보여준다. VAE와 달리 GAN에서는 3절 ⓒ의 분류 DNN(4 layers of 2D CNN)이 더 효과적이었다. Generator와 Discriminator 모두 2D CNN 레이어를 네 개를 사용했으며, 활성 함수로 둘 다 마지막 레이어는 Sigmoid를 사용했다. 그 이외 레이어에서의 활성함수로는, Generator의 경우 Leaky ReLU, Discriminator의 경우에는 Tanh를 사용했다.

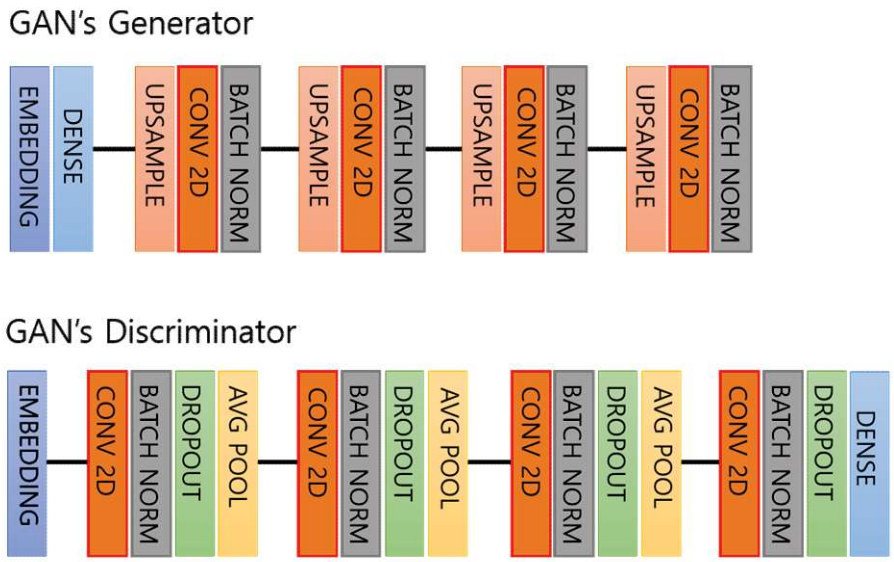


그림 8. 구축한 GAN 기반의 모델

이러한 GAN 모델을 기반으로 두 가지 ASMR 음원을 합성하는 모델을 그림 9과 같이 구축하였다. 레이블의 Embedding 결과의 평균을 새로운 레이블로 사용하고 새로운 레이블과 노이즈를 입력으로 사용하여 GAN 모델을 만들었다.

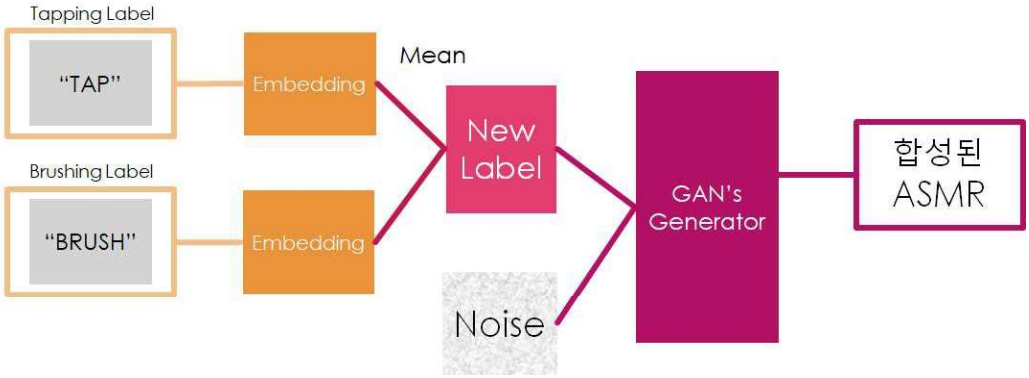


그림 9. GAN 기반의 ASMR 음원의 합성 생성모델

제 5 절 실 험 결 과

5-1. 실험 환경

우리의 실험은 Keras 프레임워크에서 진행되었다. 오디오 파형 및 샘플링 그리고 MelSpectrogram 계산을 위해서는 Librosa 라이브러리 패키지를 사용했다[5]. 제안하는 DNN의 hyperparameter 최적화를 위해서는 Talos 라이브러리를 사용했다[6]. 실험은 NVIDIA Tesla K80을 사용했고 DNN을 학습하는 과정에서 Epoch당 약 60초로, 총 30분의 학습시간이 필요했다.

5-2. 분류 DNN 모델의 정확도

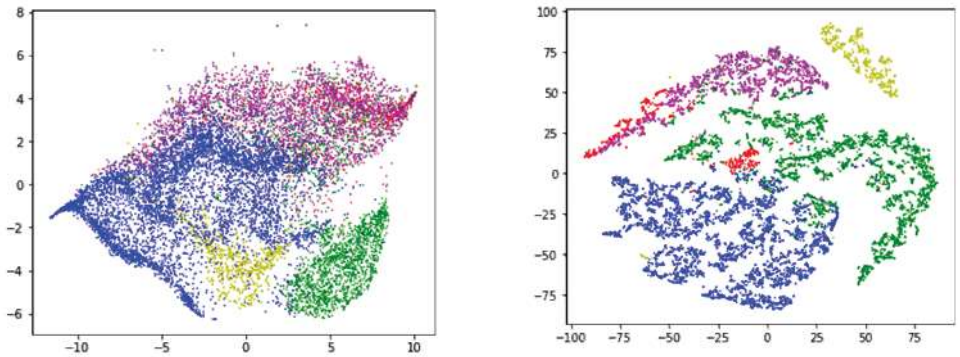
표 1은 3-2절에서 기술한 ASMR을 인식하기 위한 6가지 DNN 모델의 정확도를 보여주고 있다. 오디오 기반의 dense DNN과 음성인식의 2D CNN은 ASMR에 대해 매우 낮은 정확도를 보이는 데 이는 ASMR이 음성과 음악 등과는 다른 특성을 보이기 때문으로 보인다. 그에 비해 생활소음에 대해서 효과있던 1D CNN 모델과 이를 4개의 연결한 모델은

ASMR에서도 87%에 이르는 비교적 우수한 정확도를 얻는다. 2D CNN 레이어를 4개를 사용한 모델도 좋은 정확도를 보여주었다. 그러나 우리가 제안하는 모델이 95%의 정확도를 보여 ASMR에 최적화된 모델임을 보여주고 있다.

표1. 3-2절에서 제안한 ASMR을 인식하는 6가지 DNN 모델의 정확도

Model	Accuracy
ⓐ single dense layer on audio	40.0 %
ⓑ 2D CNN on MelSpectrogram	30.0 %
ⓒ 4 layer of 2D CNN on MelSpectrogram	85.0 %
ⓓ 1D CNN on MelSpectrogram	86.0 %
ⓔ 4 Layers of 1D CNN on MelSpectrogram	87.7 %
ⓕ 4 Layers of 1D CNN after fine tuning on MelSpectrogram	95.4 %

그림 10은 제안하는 모델 ⓕ가 오디오 기반 dense 모델보다 더 명확히 6가지의 ASMR들을 분류함을, t-Stochastic Neighbor Embedding(t-SNE) 시각화를[3] 이용해 클러스터링 결과를 보여주고 있다. 즉 더 좋은 클러스터링을 보여준다.



(a) (b)

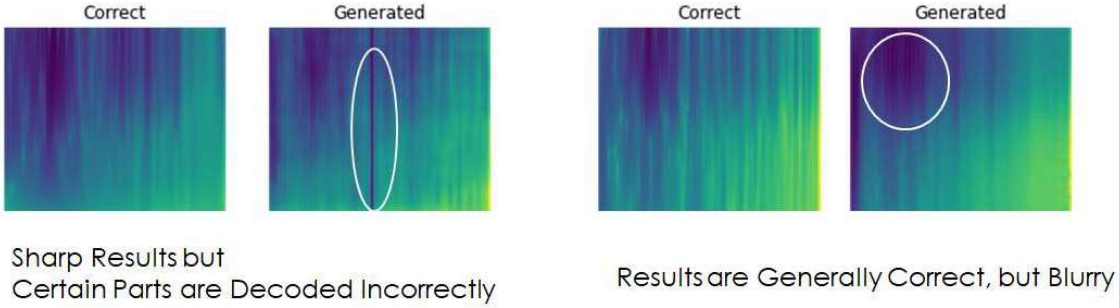
그림 10. (a)오디오기반 Dense Layer와 (b)제안하는 MelSpectogram기반 ⓕ DNN의 t-SNE 시각화 결과(ASMR 종류에 따라 다른 색으로 표현).

5-3. DNN 기반의 ASMR 생성

본 절에서는 오토인코더, VAE 및 GAN을 이용한 ASMR 생성 결과를 보고하고 평가한다.

5.3.1. 오토인코더 ASMR 생성 결과

단순 오토인코더 방식은 기존 ASMR을 변형하는 데 있어서 한계를 보여주었다. 그림 11(a)에서는 tapping 하는 ASMR에 대하여 오토인코더를 이용하여 변형된 ASMR을 생성한 결과이다. 결과는 명확했지만 디코딩이 부정확하게 되어 비정상적인 모습을 보여주는 것을 확인할 수 있다. 또한 ASMR이 제대로 변형되지 못하는 모습도 보여준다.



(a) (b)

그림 11. (a)오토인코더 방식과 (b)VAE 방식을 이용한 ASMR 생성 예

5.3.2 VAE 기반 ASMR 생성 결과

위의 예제에 대해서 VAE 방식은 그림 11(b)에서처럼 좀 더 유연한 ASMR의 변형을 수행할 수 있다. 그 이유는 러닝이 진행됨에 따라 모델이 입력의 MelSpectrogram의 분포를 추정할 수 있기 때문이다.

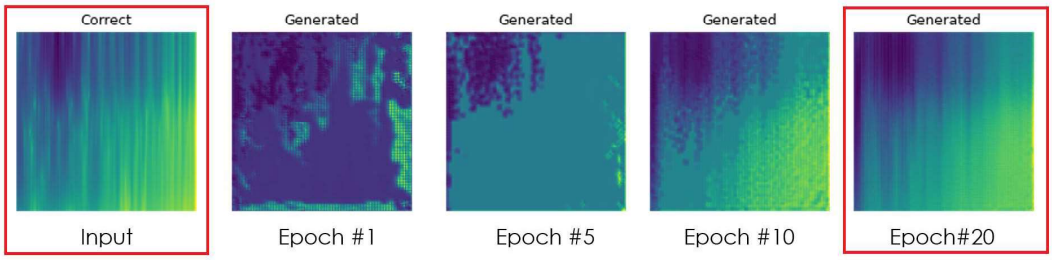


그림 12. 입력 ASMR과 학습의 Epoch 진행에 따른 ASMR 생성 결과

그림 12는 위의 ASMR에 대하여 제안한 VAE 모델을 이용하여 변형된 ASMR을 생성한 결과이다. 러닝의 Epoch이 진행함에 따라 점점 입력된 ASMR의 MelSpectrogram과 유사한 MelSpectrogram으로 변형되어 20번째 epoch에서는 매우 유사한 MelSpectrogram이 생성됨을 확인할 수 있었다.

다음은 VAE 기반으로 두 가지 ASMR 음원 합성을 통한 ASMR 생성 결과를 보인다. 그림 13-1에서는 두 가지 ASMR(Wood 두드리는 소리와 Glass 두드리는 소리)의 MelSpectrogram과 오디오 Waveform을 보여주고 있다. 그림 13-2의 왼쪽에는 이를 VAE로 합성하여 생성한 새로운 ASMR의 MelSpectrogram과 오디오 생성한 결과를 보여주고 있다. 두 소리를 각각 인코딩한 후 나온 Latent Variable을 Linear Interpolation을 한 후 디코딩하였다. 13-2의 오른쪽에는 두 소리를 인코딩 없이 단순히 Linear Interpolation한 결과를 보여준다. VAE의 결과가 두 소리의 특징을 더 반영되는 것을 확인할 수 있다. 특히 Glass 소리의 Time Domain 결과에서 파형이 가장 강하게 나타난 시간대에서의 두 합성 결과를 비교하면, 단순 Linear Interpolation 결과에서는 거의 무시되는 반면, VAE의 결과에서는 강하게 드러나는 특징이 보인다 (VAE 결과에서 붉은 박스 부분을 참고).

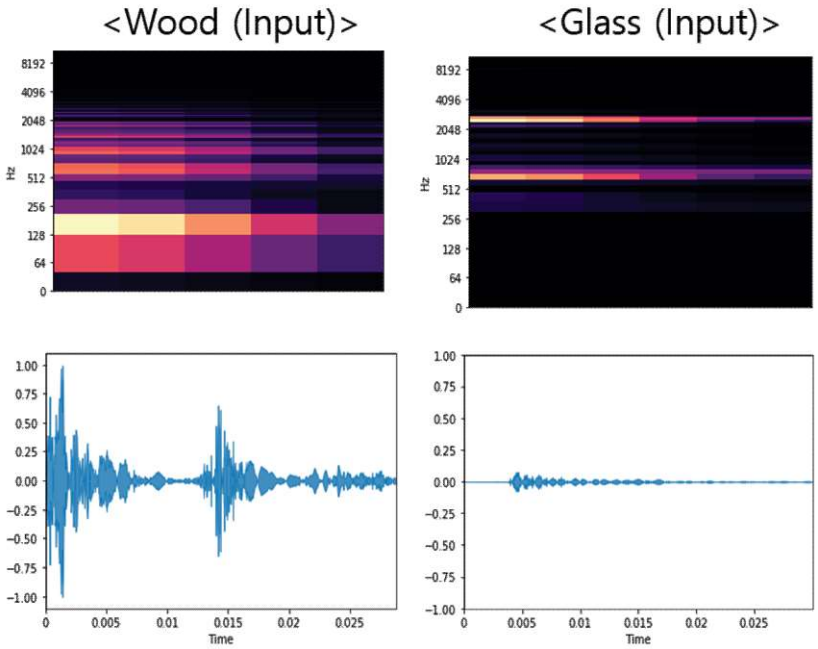


그림 13-1. 두 입력 ASMR의 MelSpectrogram과 오디오 Waveform

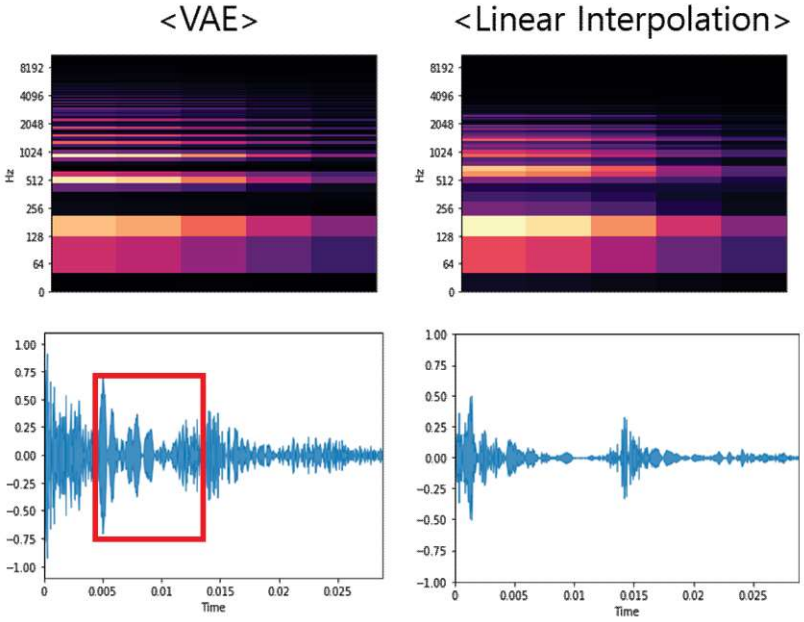


그림 13-2. ASMR 합성 결과의 MelSpectrogram과 오디오 Waveform

소리 합성이 잘 되었는지 확인하기 위해 Google Magenta[9]의 오디오 기반의 Wavenet 오토인코더를 사용하여 합성된 결과와 비교하였다. 수치적으로 비교하기 위해 사람이 인지하는 8가지 음색적 특징을 계산하는 라이브러리인 Audio Commons Timbral Model[8]을 사용한다. 표2는 이 모델에 따라 계산된 결과를 보여주는 데, Linear Interpolation은 Wood(Input)와 거의 동일한 결과를 보인 반면, VAE의 경우 Brightness 부분에서 두 소리의 중간 정도의 결과를 보인 것을 확인할 수 있다. Magenta의 경우 Brightness와 Roughness에서 중간 정도의 결과를 보였지만, Depth 또는 Sharpness의 경우 완전히 다른 결과를 보이는 것을 확인할 수 있다. 이를 통해 VAE를 사용한 합성 방식도 Linear Interpolation과 Magenta의 결과처럼 음색적인 변화를 보이지만 더 유사한 결과를 생성하는 것을 확인할 수 있다.

VAE와 Magenta의 모델을 비교하면 MelSpectrogram 기반의 VAE 모델은 오디오 기반의 Wavenet 모델에 비해 학습 시간이 매우 짧다는 장점이 있다(VAE는 MelSpectrogram을 학습하는 데 30분 이내로 완료). 그러나 MelSpectrogram을 사용한 경우 Phase 정보를 계산하지 않아 Wavenet 모델을 사용한 결과에 비해 소음이 많다는 단점이 있다. 그러나 매우 조용한 소리를 사용하는 ASMR의 경우, Wavenet 모델을 사용하여도 노이즈가 많이 발생하기 때문에 두 경우 모두 추가적인 디노이징이 필요하다. 종합적으로 보면 VAE 방식의 ASMR 음원 생성이 Wavenet 방식보다 더 효율적일 가능성이 높다고 본다.

표 2. 입력과 합성된 소리의 분석 결과 및 Magenta와의 비교

	hardness	depth	brightness	roughness	warmth	sharpness	boominess
Glass (input)	77	38	66	62	40	61	26
Wood (input)	75	38	77	76	39	61	25
My VAE	84	38	73	77	34	61	19
Linear	75	38	77	76	39	61	25
Magenta	60	34	67	71	47	58	21

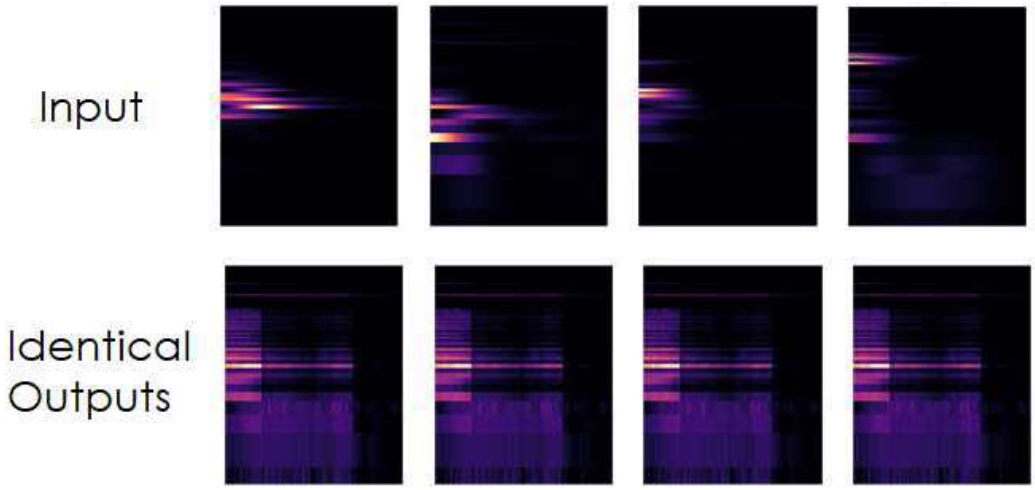


그림 14. VAE의 Posterior Collapse 문제로 인해 서로 다른 입력 ASMR에 대해서 모두 거의 동일한 변형 ASMR을 생성하는 예

VAE 생성모델에 있어서 한가지 문제점은 기존 VAE의 이슈 중의 하나로 알려진 Posterior Collapse 문제를 ASMR 생성에서도 발견할 수 있었다는 점이다[17]. 즉 러닝이 진행됨에 따라 디코더가 샘플링을 무시하여 서로 다른 ASMR 입력에 대해서도 동일한 ASMR이 생성되는 경우가 있음을 확인할 수 있었다. 이는 학습이 부진했거나 입력 노이즈 등의 영향으로 보인다. 그림 14는 네 가지 서로 다른 ASMR 입력들에 대해서

같은 ASMR 출력을 보이는 VAE 학습 케이스를 보여준다.

5.3.3. GAN 기반 ASMR 생성 결과

그림 15는 주어진 ASMR에 대하여 제안한 GAN 모델을 이용하여 변형된 ASMR을 생성한 결과이다. 동일한 레이블에 대한 Generator 결과를 입력과 비교하면 학습이 진행함에 따라 점점 입력된 ASMR의 MelSpectrogram과 유사한 MelSpectrogram으로 변형되어 18번째 epoch에서는 매우 유사한 MelSpectrogram이 생성됨을 확인할 수 있었다.

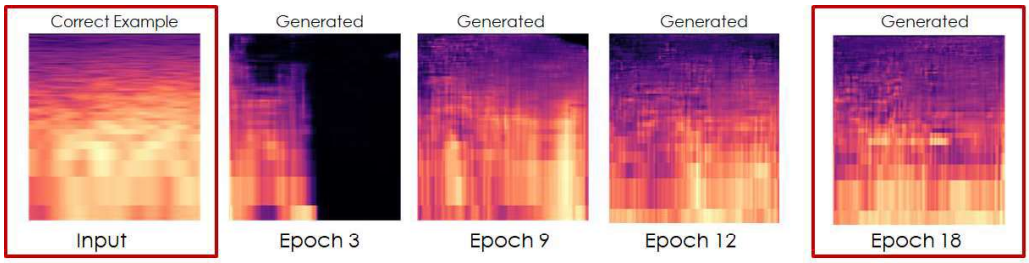


그림 15. GAN을 이용해 2가지 ASMR 합성 결과 MelSpectrogram

5-4. VAE 생성모델과 GAN 생성모델 정확도 비교

본 절에서는 VAE 생성모델과 GAN 생성모델의 ASMR 생성성능을 비교하고자 한다. 첫 번째로는 우선 주어진 ASMR 음원들에 대해 변형과 합성을 통해 생성된 ASMR 음원의 MelSpectrogram 모습과 실제 소리의 데모를 비교한다. 두 번째로는 랜덤하게 생성된 MelSpectrogram들을 5-2절의 분류 DNN 모델로 분류하여 정확성을 평가한다.

5.4.1. 생성된 ASMR MelSpectogram의 비교

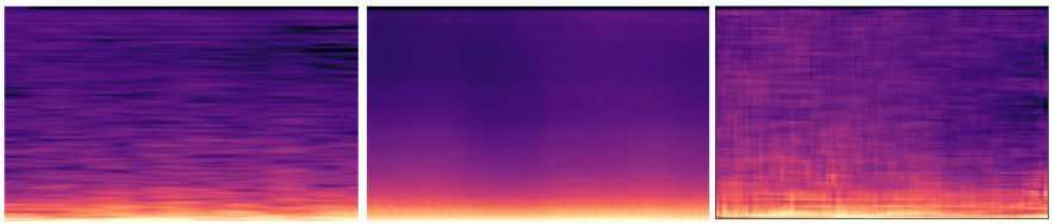
그림 16은 두드리는 ASMR 소리에 대한 원본 MelSpectrogram과 VAE

로 생성한 MelSpectrogram, 그리고 GAN이 생성한 MelSpectrogram을 보여준다. VAE와 GAN 모두 유사하나 다른게 변형된 ASMR을 생성함을 확인할 수 있다.

(a) (b) (c)

그림 16. (a)원본 ASMR (b)VAE 생성 ASMR (c) GAN 생성 ASMR

그림 17은 솔질하는 ASMR 소리에 대한 원본 MelSpectrogram과 VAE로 생성한 MelSpectrogram, 그리고 GAN이 생성한 MelSpectrogram을 보여준다. VAE와 GAN 모두 유사하나 다르게 변형된 ASMR을 생성함을 확인할 수 있다. 그림 16과 그림 17 ASMR 소리에 대한 데모는 [16]에서 확인할 수 있다.



(a) (b) (c)

그림 17. (a)원본 ASMR (b) VAE 생성 ASMR (c) GAN 생성 ASMR

그림 18은 그림 16의 두드리는 소리와 그림 17을 솔질 소리를 VAE와 GAN을 이용하여 합성한 소리를 보여주고 있다. 그림 18은 VAE와 GAN은 각각 타원으로 표시된 영역에서 두 소리의 특징이 존재하면서도 다른 소리 결과가 적절히 합성되었음을 보여주고 있다.

MelSpectrogram을 시각으로 확인했을 때에는 GAN의 결과가 더 합성을

선명히 보여주고 있지만 다음 절에서 보듯이 분류 DNN은 다르게 판단하는 것으로 보인다.

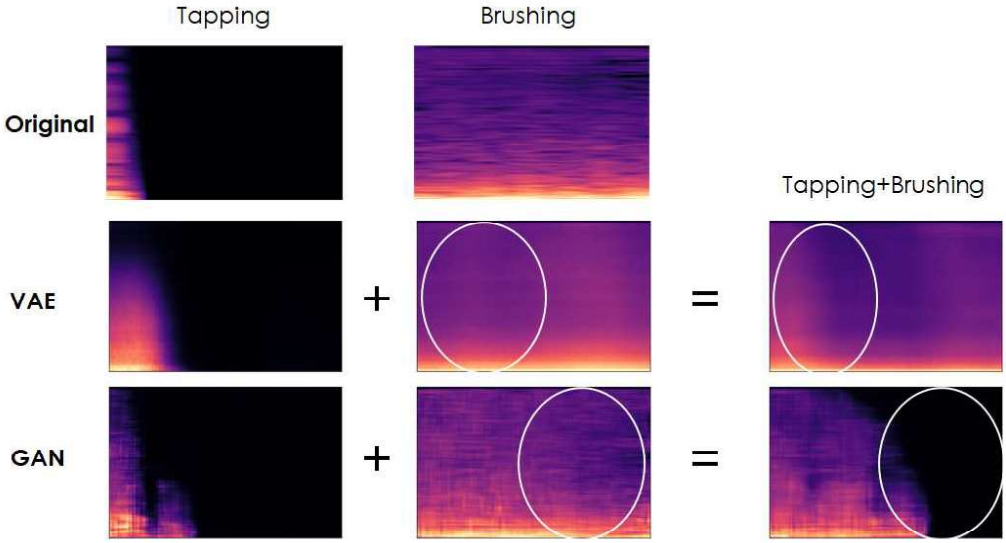


그림 18. GAN을 이용해 2가지 ASMR 합성 결과 MelSpectrogram

5.4.2. 분류 DNN을 이용한 생성된 ASMR의 비교

VAE와 GAN을 이용하여 생성한 ASMR 음원을 우리의 분류 DNN에 입력으로 넣어 그 정확도를 평가하였다. 이를 위해 변형된 소리의 경우 레이블 당200개의 샘플을 생성하였고 두드리는 소리와 솔질 소리를 합성한 소리의 경우 100개의 샘플을 생성하였다. 정확도는 변형된 소리의 경우 원래 음원으로 분류되는 확률을 측정하고 합성된 소리의 경우 두 소리로 분류하는 확률의 합이 50%를 넘으면 정확한 것으로 측정되었다. 표3은 변형된 소리와 합성된 소리의 정확도를 보여준다. 변형된 소리의 경우 70% 혹은 65% 이상이 성공적으로 통과하였고 합성된 소리는 둘 중 하나만 통과하면 되므로 더 높은 정확도를 보였다. 이 결과는 비교적 생성된 소리가 유사함을 보여준다. 그리고 시각적으로 보이는 것과는 다르게 VAE가 GAN 보다 약간 더 높은 정확도를 보여주었다.

본 논문에 대한 코드 및 자세한 사항은 Github에 공개되어있다.[18]

표 3. 변형 및 합성된 소리의 정확도

	변형된 소리	합성된 소리
VAE	71 %	93 %
GAN	64 %	86 %

제 6 절 요약 및 향후 과제

본 논문에서는 최초로 딥러닝 기반의 ASMR 개발 플랫폼을 제안하였고 ASMR 인식을 위한 DNN 모델을 구축하였다. 제안한 모델은 실제 ASMR들의 인식에 매우 탁월한 정확도를 보였다.

또한 본 연구에서는 GAN 구조와 VAE 구조를 이용하여 ASMR을 변형하고 합성하여 새로운 ASMR을 생성하는 모델을 제안하였다. 제안하는 모델을 통해 하나의 ASMR의 변형과 서로 다른 ASMR의 합성이 비교적 효과적으로 이루어짐을 실험을 통해 확인하였다. 이러한 결과는 ASMR 음원의 생산성을 높일 가능성을 보여주었다.

추후에 ASMR 데이터베이스에 더 많은 종류의 소리를 추가하고, 이 모델을 기반으로 레이블이 불분명한 소리에 대해서도 분류하여 데이터베이스에 추가하는 작업이 필요하다. 또한 더 많은 ASMR에 대하여 생성 실험을 진행하여 모델의 완성도를 높이는 작업이 필요하다. 오디오 기반의 모델과의 비교하여 MelSpectrogram 방식이 러닝 시간뿐만 아니라 생성된 ASMR 결과의 질적인 면에서도 우수한지를 평가하는 것이 흥미로운 추후 과제이다. 더불어 생성된 ASMR을 효과적으로 디노이징하는 방법을 모색하여 더욱 우수한 품질의 ASMR를 생성하는 것도 필요하다.

참 고 문 헌

[1] Barratt, Emma L, and Nick J Davis. “Autonomous Sensory Meridian Response (ASMR): a flow-like mental state.” PeerJ vol. 3 e851. doi:10.7717/peerj.851 (2019)

[2] Abdoli, Sajjad et al., “End-to-End Environmental Sound Classification using 1D Convolutional Neural Network” (2019)

[3] Laurens van der Maaten, “Learning a Parametric Embedding by Preserving Local Structure”, 12th International Conference on Artificial Intelligence and Statistics, (2009).

[4] Purwins, Hendrik et al. “Deep Learning for Audio Signal Processing.” IEEE Journal of Selected Topics in Signal Processing 13.2 (2019)

[5] LibROSA, https://librosa.github.io/librosa/

[6] Talos, https://github.com/autonomio/talos

[7] Wenhao bian et. al, “Audio-based music classification with DenseNet and data augmentation, PacificRIM International Conference on AI, Fiji (2019).

[8] K. Choi, G. Fazekas, and M. Sandler.: Automatic tagging using deep convolutional neural networks. Society for Music Information Retrieval Conf., NY (2016).

[9] Small sounds, big money: The commercialization of ASMR, https://www.wired.com/story/commercialization-of-asmr/

[10] An Introduction to Variational Autoencoders
Foundations and Trends® in Machine Learning, 2019, Kingma, Diederik P. and Welling, Max

[11] Goodfellow et. al, Generative Adversarial Networks, 2014 Annual Conference on Neural Information Processing Systems(NIPS), 2014.

[12] Audio Commons Timbral Model

https://github.com/AudioCommons/timbral_models
D5.8: Release of timbral characterisation tools for semantically annotating non-musical content, January 2019. License: CC-BY 4.0.
[13] J. Engel et. al, Neural Audio Synthesis of Musical Notes with Wavenet Autoencoders. CoRR, abs/1704.01279, 2017.
[14] "Oddly IKEA": IKEA ASMR, https://www.youtube.com/watch?v=uLFaj3Z_tWw&feature=emb_title
[15] M. Mirza and S. Osindero, Conditional Generative Adversarial Nets, 2014
[16] Demo of ASMR generation by variating ASMRs using VAE and GAN, Youtube address. https://youtu.be/v9ND1fJYLpk
[17] Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick, Lagging Inference Networks and Posterior Collapse in Variational Autoencoders, 2019
[18] DeepASMR Github Link: https://github.com/moonjee/DeepASMR

Abstract

DeepASMR: Deep Learning-based ASMR Platform

Jee Young Moon
Electrical and Computer Engineering
The Graduate School
Seoul National University

Recently, there has been growing interests for "special" sounds that can trigger psychological comfort, called ASMR (Autonomous Sensory Meridian Response), and related businesses have already started to bloom. However, there exists the low productivity issue for ASMR sound creation as the process of creation takes a long time with tremendous amount of effort. This dissertation proposes *DeepASMR*, an ASMR platform based on deep learning to tackle this issue. DeepASMR uses DNN to collect and classify ASMR sounds, and to create new ASMR sounds based on existing ones. Our proposed DNN model can increase the accuracy of classification of ASMR sounds up to 95%, surpassing the existing DNN models used for noise or music classification. This makes it efficient to collect ASMR sounds from the internet and create a database, thus letting users choose their ASMR trigger sounds easily. We can also generate new ASMR sounds using DNN by modifying or merging trigger sounds. We

constructed the generation DNNs using the VAE (Variational Autoencoder) model and the GAN (Generative Adversarial Network) model. When these newly-created ASMR sounds are given to our classification DNN as inputs, the accuracy was over 70%, indicating that the proposed DNNs generated high-quality ASMR sounds.

keywords : ASMR, Deep learning, classification, generation, VAE, GAN
Student Number : **2019-24791**