오디오 소스 분리 시 발생되는 아티팩트를 이용한 딥페이크 노래 딥러닝 탐지 모델

서영민
㈜브레인데크

ym@braindeck.net

Detecting DeepFake Songs using Artifacts in Audio Source Separation

Seo Youngmin
Braindeck Inc.

요 약

본 연구는 딥페이크 노래 탐지를 위한 새로운 방법론을 제안한다. 오디오 소스 분리 과정에서 발생하는 아티팩트를 이용하여 딥페이크를 식별하는 이 방법은 기존의 보코더 기반 탐지 방식과 차별화된다. Musdb18HQ 데이터셋을 사용하여 실험을 진행하였으며, 원본 목소리와 변환된 목소리의 멜스펙트로그램을 CNN 모델로 분류하였다. 실험 결과, 높은 정확도를 보였으나 F1 스코어가 상대적으로 낮게 나타났다. 이는 임계 값 최적화, 장르별 보컬 특성 차이 등의 문제로 인한 것으로 추정된다. 그럼에도 본 논문의 실험 결과는 오디오 소스 분리에서 발생되는 아티팩트를 이용한 탐지 방법이 딥페이크 노래 탐지에 매우 강력한 증거를 제공할 수 있는 가능성을 시사한다.

Ⅰ. 서 론

인공지능 기술이 발전함에 따라 이를 이용한 딥페이크 기술 또한 나날이 정교해지고 있다. 최근 글로벌 스트리밍 플랫폼 스포티파이(Spotify)에 미국의 유명 가수 드레이크(Drake)와 더 위켄드(The Weeknd)의 목소리를 이용한 인공지능 생성 곡이 상당한 스트리밍 수와 조회수를 기록한 사건은 이 기술의 잠재적인 위험성을 여실히 보여주었다. 이는 장난이나 단순한 사기의 차원을 넘어 심각한 저작권 침해 문제를 야기할 수 있음을 보여준다.

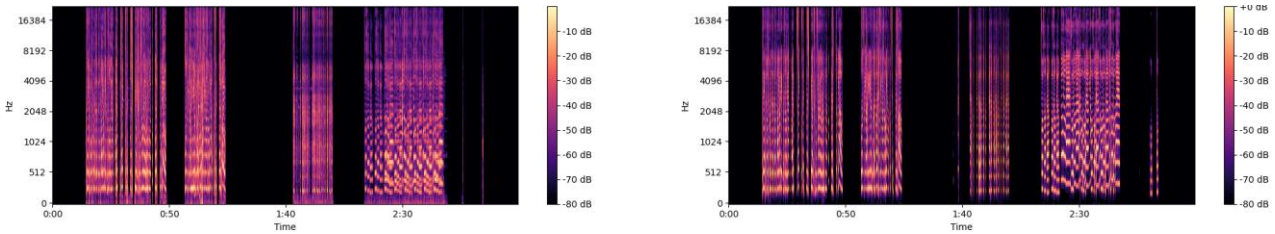
그러나 음성 관련된 딥페이크 탐지에 대한 연구는 여전히 제한적이며, 음악 산업에서의 저작권 보호를 위한 노래 목소리의 딥페이크 탐지 방법론은 미흡한 실정이다. 현재 음성 딥페이크 탐지 방법론 대부분은 보코더(Vocoder) 탐지에 집중되어 있다[1]. 보코더의 아티팩트(Artifact)를 이용한 탐지 방법은 유의미한 성과를 보이고 있으나, 보코더를 사용하지 않는 최신 End-to-End 딥러닝 모델에는[2], [3] 적용하기 어렵다는 한계점이 존재한다. 또한 기존 연구 대부분이 노래가 아닌 짧은 대화 데이터를 쓰고 있어, 노래의 딥페이크 탐지에는 부적합하다.

본 연구에서는 이러한 한계를 극복하고자 오디오 소스 분리에서 발생하는 아티팩트를 이용한 새로운 접근 방식을 제시한다. 딥페이크 노래 제작 과정에서 필수적으로 사용되고 있는 오디오 소스 분리(Audio Source Separation) 기술에 착안해, 이 과정에서 발생하는 아티팩트를 탐지함으로써 딥페이크 노래를 식별하는 방법을 제안한다. 이 방법은 기존의 보코더 기반 탐지 방식과 차별화될 뿐만 아니라, 다양한 딥페이크 생성 기술에 대해 보다 범용적으로 적용할 수 있는 가능성을 제시한다.

Ⅱ. 본론

본 실험은 오디오 소스 분리 시 아티팩트가 발생하는 것을 발견하기 위해 음성 음원과 악기 음원이 분리된 데이터셋인 Musdb18HQ(인용)으로 진행하였다. 해당 데이터셋은 44.1KHz 의 샘플링 레이트 가진 고품질 노래 데이터셋으로 100 개의 훈련용 노래와 50 개의 테스트용 노래로 구성되어 있으며, 각 노래는 각기 다른 길이와 악기 구성을 갖는다. 모든 곡들은 보컬 트랙과 각 악기 트랙으로 분리되어 있다.

오디오 소스 분리에는 Band-Split RoPE Transformer[4], Mel-Band RoFormer[5], 파인튜닝 된 MDX-Net[6] 3 가지 모델을 사용해 음원 분리를 수행했으며, 분리된 음원을 VITS (Variational Inference Text-to-Speech)[3] 와 HiFi-GAN[7]을 사용해 분리된 보컬 트랙을 미국의 유명 여가수 레이디 가가(Lady Gaga)의 목소리를 학습한 모델을 이용해 딥페이크 버전을 생성하였다. 해당 모델은 레이디 가가의 음원 50 곡을 이용했고, 500 번의 에폭(epoch)으로 설정해 학습하였다. 각 샘플은 멜스펙트로그램으로 변환 후 윈도우의 크기는 2048, 홉 길이는 512 로 설정해



(a) 원본 목소리 멜스펙트로그램 (b) 변환된 목소리 멜스펙트로그램

그림 1 원본 목소리와 변환된 목소리의 멜스펙트로그램 예시

프레임 단위로 분할하였다. 데이터셋의 구성은 표 1 과 같다.

표 1 데이터셋 구성

데이터셋	원본 목소리	변환된 목소리
훈련 데이터 수	40	40
검증 데이터 개수	10	10
테스트 데이터 수	25	25
훈련 데이터 프레임 수	1,579	1,547
검증 데이터 프레임 수	375	407
테스트 데이터 프레임 수	1,063	1,063

그림 1 의 (b)를 (a)와 비교하면, 변환된 목소리는 저음 주파수 부분이 나타나지 않는다. 이는 모든 변환된 곡에서 일괄되게 나타나는 특징이다. 또한 원본 목소리에 비해 변환된 목소리의 데이터 분포를 살펴보면 공백 값을 가지는 경우가 더 많으며, 파워의 분포도 고르지 않은 샘플이 많은 것을 확인할 수 있었다.

본 실험에서는 오디오 소스 분리 중 모델의 필터링 과정에서 손실이 발생한다는 가정과 이것이 딥페이크 노래 탐지의 아티팩트로 적용될 수 있을 것이라 생각해 CNN 모델을 이용해 멜스펙트로그램을 이진 분류하는 실험을 진행했다. 분류에 사용한 모델은 ResNet50, EfficientNetV2, DenseNet169 로 3 가지다. 각 모델은 IMAGENET 으로 사전학습 된 가중치를 사용했으며, 옵티마이저는 Adam, 학습률은 0.001, 배치 사이즈는 128, 에폭은 200 으로 설정했으며 조기 종료를 사용해 학습을 진행했다. 실험 장비로는 NVIDIA A100 80GB GPU 1 개를 사용하였다. 실험은 모두 5 번을 진행했으며 평균 값을 최종 결과로 사용했다.

실험의 결과 값은 표 2 와 같다. EffcientNetV2 가 가장 높은 정확도를 보였으나 다른 모델과 정확도 차이는 약 0.1 정도로 그렇게 크지 않는 것으로 나타났다. 각 모델의 결과는 전체적인 성능은 높은 정확도를 기록했으나 정확도와 비교해 낮은 F1 스코어를 나타냈다.

표 2 실험 결과

모델	정확도	F1 스코어
ResNet50	0.9487±0.0099	0.5194±0.0053
EfficientNetV2	0.9552±0.0036	0.5219±0.0016
DenseNet169	0.9431±0.0099	0.5176±0.0049

F1 스코어가 정확도에 비해 낮은 이유는 임계 값 최적화 부재, 장르별 보컬의 특이성에 따른 특성 차이, 노래의 무음 부분 분류 시 정확한 클래스 분류를 하지 못했을 것으로 보인다. 또한 CNN 모델의 시간적인 특성 학습에 한계점이 존재해 이러한 결과에 영향을 미쳤을 수 있다. F1 스코어를 개선하기 위해 임계 값의 최적화 진행과 노래의 장르별 레이블링, RNN, Transformer 레이어를 추가하는 방법을 사용해볼 수 있다.

Ⅲ. 결론

본 연구에서는 오디오 소스 분리 과정에서 발생하는 아티팩트를 이용하여 딥페이크 노래를 탐지하는 새로운 방법을 제안하고 실험을 통해 그 가능성을 확인하였다. 실험 결과, 제안된 방법은 높은 정확도를 보여주었으나, F1 스코어가 상대적으로 낮게 나타나는 한계점도 발견되었다. 이러한 결과는 오디오 소스 분리의 아티팩트가 딥페이크 노래 탐지에 유용한 특징으로 활용될 수 있음을 시사한다. 그러나 동시에 임계 값 최적화, 장르별 보컬 특성 차이, 무음 부분 처리 등의 문제를 해결할 필요성도 제기되었다.

향후 연구에서는 한계점을 극복하기 위해 장르별 노래에 대한 레이블링 및 데이터셋, 다수의 목소리를 이용한 실험 설계, 시간적 특성을 고려할 수 있는 모델 도입을 기반으로 연구를 확장할 계획이다. 추가적인 실험을 통해 딥페이크 노래 탐지의 정확도와 신뢰성을 더욱 향상시키고, 나아가 음악 산업에서의 저작권 보호와 AI 윤리 문제 해결에 기여할 수 있을 것으로 기대된다.

참 고 문 헌

[1] R. Du, J. Yao, Q. Kong, and Y. Cao, “Towards Out-of-Distribution Detection in Vocoder Recognition via Latent Feature Reconstruction,” Jun. 04, 2024

[2] D. Diatlova and V. Shutov, “EmoSpeech: Guiding FastSpeech2 Towards Emotional Text to Speech,” Jun. 28, 2023

[3] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *International Conference on Machine Learning*, PMLR, 2021, pp. 5530–5540.

[4] W.-T. Lu, J.-C. Wang, Q. Kong, and Y.-N. Hung, “Music Source Separation With Band-Split Rope Transformer,” in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of: IEEE, Apr. 2024, pp. 481–485

[5] J.-C. Wang, W.-T. Lu, and M. Won, “Mel-Band RoFormer for Music Source Separation,” arXiv.org. Accessed: Jul. 16, 2024

[6] “KUIELab-MDX-Net: A Two-Stream Neural Network for Music Demixing,” *arXiv.org*, Nov. 2021

[7] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis,” Oct. 23, 2020