연성 워터마킹 기반 오디오 딥페이크 탐지

김준모, 한창희
서울과학기술대학교 전기정보공학과

0904junmong@seoultech.ac.kr, chahn@seoultech.ac.kr

Deepfake Detection with Audio Fragile Watermarking

Jun-Mo Kim, Changhee Hahn
Dept. of Electrical an Information Engineering, Seoul Tech

요 약

디지털 오디오 파일의 보안은 디지털 미디어의 확산과 함께 점차 중요해지고 있다. 특히, 딥페이크와 같은 기술을 이용한 조작이 증가함에 따라, 이를 효과적으로 방지하는 기술이 대두되고 있다. 본 연구에서는 연성 워터마킹 기술을 활용하여, 오디오 파일이 외부 조작에 의해 변경되었을 때 오디오 파일이 의도적으로 파괴하는 방식을 제안한다. 본 논문에서는 연성 워터마크 생성 및 삽입 방법에 관한 자세한 설명을 하고, 연성 워터마킹을 통해 오디오의 변조 여부를 즉각적으로 탐지하는데 어떻게 기여하는지를 보여준다. 제안 기법은 오디오 원본의 무결성을 효과적으로 보호하는 새로운 방법을 제시하며, 디지털 미디어 보안을 강화하는데 중요한 역할을 할 것으로 기대된다.

1. 서론

최근 인공지능의 급속한 발전은 많은 긍정적인 변화를 가져왔지만, 동시에 디지털 미디어의 조작의 위험도도 증가시켰다. 특히, 딥페이크 기술은 정치적 목적으로 허위 정보를 확산하거나, 유명 인사들을 대상으로 한 사기와 명예 훼손 등의 사회적 문제를 야기한다. 이러한 문제에 효과적으로 대응하기 위해서는 디지털 콘텐츠의 변조 여부를 신속하게 판별할 수 있는 효과적인 방법이 필수적이다.

오디오 워터마킹은 오디오 신호에 보이지 않는 정보를 삽입하여 그 원본의 무결성을 보장하는 기술이다. 본 논문에서는 워터마크에 조작이 가해졌을 때 오디오의 내용을 파괴함으로써 원본이 아님을 즉각적으로 알 수 있게 하는 연성 워터마킹 기술을 사용한다. 이전의 워터마킹 방식들이 워터마크의 파괴 여부만을 확인하였다면, 본 연구에서는 수정이나 변조가 감지될 때 청각적으로 인지할 수 있을 정도로 오디오의 변조를 탐지하는 새로운 기법을 제시한다.

본 논문은 이러한 연성 워터마킹의 기술적 구현과 그 효용성을 다루면서, 디지털 미디어의 보안을 강화하는 새로운 방안을 모색한다.

2. 연성 워터마킹

오디오 워터마킹은 오디오 파일 내에 삽입된 워터마크가 오디오의 수정 또는 변조가 발생 시 손상되거나 파괴되는 특성을 가진다. 이러한 워터마크의 취약성은 의도적으로 설계된 것으로, 오디오 파일의 무결성을 보호하고, 조작이 시도되었을 경우 즉각적으로 탐지할 수 있게 한다. 다음은 본 연구에서 사용된 연성 워터마킹 기술의 원리에 대해 설명한다.

2-1) 워터마크 생성

연성 워터마크는 난수 생성기로부터 얻은 기본키를 바탕으로 생성이 된다. 기본 키에 특정 문자열을 추가한 후, 이를 암호 해시 함수(예, SHA-256)로 처리하며[1], 생성된 해시 값을 워터마크로 활용된다. 특히, 워터마크는 이진 형태의 데이터로 변환되어 각 비트가 ‘1’ 또는 ‘0’ 상태로 오디오 샘플에 적용된다.

2-2) 워터마크 삽입

연성 워터마크의 삽입은 오디오 파일의 시간 도메인에서 직접 수행된다. 워터마크의 각 비트에 따라 오디오 샘플의 값을 미세하게 조정하는데, 예를 들어 비트가 ‘1’인 경우 해당 오디오 샘플의 진폭을 소폭 증가시키고, ‘0’인 경우 감소시킨다. 이러한 조정은 매우 미세하게 이루어져 일반적인 청취 상황에서는 인지할 수 없다. 그러나 이 조정은 오디오의 어떤 변조나 수정에 의해 쉽게 파괴될 수 있는 특성을 지닌다.

3. 무결성 평가 기준

3-1) 신호 대 잡음비(SNR, Signal Noise Ratio)

연성 워터마킹 기술의 무결성 검증을 위해 신호

대 잡음비(SNR)를 이용한다. SNR 은 원본 오디오 신호의 파워 대비 워터마크 삽입으로 인한 잡음 파워의 비율을 데시벨 단위로 나타낸다. 이 지표는 워터마크가 오디오에 얼마나 미세하게 삽입되었는지를 정량적으로 평가하는 데 사용된다.

3-2) 유클리디안 거리

유클리디안 거리는 두 점 사이의 직선 거리를 계산하는 가장 기본적인 거리 측정 방법이다. 이 거리는 다차원 공간에서 두 데이터 포인트 간의 차이를 정량적으로 나타내는데 사용된다. 오디오 데이터 분석에서 유클리디안 거리는 두 오디오 샘플 간의 차이를 측정하는데 사용될 수 있다. 측정된 차이를 통해서 오디오의 파괴된 정도를 측정할 수 있다.

4. 실험

본 연구는 Librispeech 의 100 개 오디오 데이터셋을 선별하여 실험을 진행하였다[2]. 데이터셋 선별의 기준은 다양성을 위해 다양한 연령대와 성별 인종을 포함하고 있으며, 연설이나 일상 대화, 오디오 북의 내용이 주로 포함되어 있다.

본 실험에서는 각각의 오디오 파일을 읽고 디지털 신호를 시간 도메인에서 다루는데, 무작위로 생성한 키 값을 해시 함수를 통해 연성 워터마크를 생성했다. 이후 연성 워터마크를 시간 도메인의 파형에 삽입했다.

워터마크가 삽입된 오디오 파일의 무결성을 측정하기 위해 SNR 을 사용했다. <표 1>은 본 연구의 연성 워터마크와 비교군 워터마크들의 SNR 비교를 나타내고 있다 [3].

Watermark	SNR
연성 워터마크	38.70dB
WavMark [3]	36.85dB
Audiowmark [5]	35.28dB

<표 1> 타 워터마크와의 SNR 비교

위 데이터셋에 대하여 SNR 을 측정하였을 때 평균 38.70dB 이 측정되었다. 이는 다른 워터마킹 기법인 WavMark 대비 차이가 거의 없음을 감안하였을 때, 이 연구에서 사용된 방식이 무결성을 보임을 확인할 수 있다.

또한, 본 실험은 연성 워터마크가 삽입된 오디오와 워터마크가 삽입되지 않은 원본 오디오에 대해 딥페이크를 진행했다. 연성 워터마크가 삽입된 오디오에 대해서 파괴된 정도를 측정하기 위해 SNR 과 유클리디안 거리를 이용한다. <표 2>는 파괴된 오디오 파일과 노이즈 환경의 SNR 을 비교하여 오디오 파일의 손상된 정도를 보여주고 있다 [4].

Audio	SNR
연성 워터마크 적용 오디오	-1.700dB
LA-VocE [4]	-3.319dB

<표 2> 노이즈 환경과의 SNR 비교

연성 워터마크가 삽입된 오디오의 SNR 을 측정하였을 때 평균 -1.7dB 로 측정되었다. 이는 신호의 전력이 잡음의 전력보다 작아 노이즈 환경과 유사하다는 것을 의미한다. 유클리디안 거리는 원본 오디오에 대한 딥페이크 파일과 연성 워터마크가 삽입된 오디오에 대한 딥페이크 파일 사이에서 측정하였다. 데이터 셋들로 실험을 진행한 결과 유클리디안 거리의 평균값은 약 1437 인데, 이는 데이터 간 유사할 경우 유클리디안 커리가 0 에 가까워지는 특성을 고려해 볼 때 연성 워터마크가 삽입된 오디오와 원본 오디오가 매우 상이함을 보여준다.

5. 결론

본 연구에서는 디지털 오디오의 보안을 강화하기 위한 연성 워터마킹 기술의 구현과 그 효과를 검토하였다. 연성 워터마킹 기술은 오디오 파일에 삽입된 워터마크가 조작이 시도될 때 쉽게 손상되는 설계이다. 본 연구를 통해 연성 워터마크를 적용한 오디오 파일은 조작 시 높은 수준의 탐지 가능성을 보여주었으며, 신호 대 잡음비(SNR)과 유클리디안 거리를 사용한 실험 결과는 이 기술의 유효성을 확인시켜 준다.

이와 같은 결과는 디지털 오디오의 보호에 연성 워터마킹이 매우 유용한 도구임을 시사하며 오디오뿐만 아니라 다른 미디어에도 적용 가능한 범용적인 보안 솔루션으로 확장될 수 있을 것으로 기대된다. 향후 연구에서는 다양한 공격 시나리오와 환경에서의 워터마킹 기술의 효율성을 발전시킬 필요성이 있다.

참고문헌

[1] ALSABHANY, Ahmed A.; ALI, Ahmed Hussain; ALSAADI, Mahmood. A lightweight fragile audio watermarking method using nested hashes for self-authentication and tamper-proof. *Multimedia Tools and Applications*, 2024, 1-15.

[2] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in ICASSP, 2015.

[3] CHEN, Guangyu, et al. Wavmark: Watermarking for audio generation. *arXiv preprint arXiv:2308.12770*, 2023.

[4] MIRA, Rodrigo, et al. LA-VocE: Low-SNR audio-visual speech enhancement using neural vocoders. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023. p. 1-5.

[5] Stefan Westerfeld, “Audiowmark: Audio watermarking,” https://uplex.de/audiowmark, 2020.