짧은 발화 기반 딥보이스 탐지기의 코덱 강건성 분석

최길한, 김민선, 명서아, 김용강*
국립공주대학교
ygkim@kongju.ac.kr

Evaluating Codec-Induced Degradation in Short-Utterance Deepfake Speech Detection

Gil Han Choi, Min Seon Kim, Seo Ah Myeong, Yonggang Kim*
Kongju National University.

요 약

본 논문은 2초 내외의 짧은 발화 환경에서 AASIST-L 기반 딥페이크 음성 탐지기의 코덱 및 채널 변형에 대한 강건성을 분석하였다. FoR 데이터셋의 for-2sec(clean)로 학습한 모델을 동일 test 음원에 대해 AAC/MP3/Opus 트랜스코딩(codec-sim) 조건과 재녹음 기반 채널(for-rerec) 조건으로 평가하여, 코덱 종류 및 비트레이트에 따른 성능 민감도 차이를 정량적으로 비교하였다. 실험 결과, 일부 코덱 조건에서는 clean 대비 성능 변화가 제한적이었으나 Opus 계열에서 비트레이트에 따른 성능 저하가 관측되었고, 재녹음 기반 채널 조건에서는 코덱 변형보다 훨씬 큰 성능 저하(EER 증가)가 발생하였다. 이는 실제 통신 환경을 고려한 강건성 학습 및 평가의 필요성을 시사한다.

Ⅰ. 서 론

최근 딥러닝 기술의 발전으로 생성된 고품질의 합성 음성은 실제 화자의 목소리와 구분이 어려울 만큼 정교해지고 있다. 이러한 기술은 콘텐츠 제작 등 긍정적인 측면이 있으나, 보이스피싱 및 음성 인증 우회 등 악의적인 목소리 복제(Deepfake speech) 범죄에 악용될 우려가 커지며 사회적 위협으로 대두되고 있다. 이에 대응하여 최근에는 AASIST[1]와 같은 그래프 어텐션 기반의 딥보이스 탐지 모델이 제안되어 우수한 성능을 입증해 왔다.

그러나 기존의 탐지 모델들은 주로 정제된(Clean) 오디오 데이터를 기반으로 학습되어, 실제 서비스 환경에서의 강건성이 부족하다는 한계가 있다. 실제 카카오톡, 디스코드와 같은 VoIP 통화 환경에서는 전송 효율을 위해 Opus, AAC, MP3와 같은 손실 코덱(Lossy Audio Codec) 압축을 거치게 된다. 특히 실제 시나리오에서는 수 초 이내의 짧은 발화(Short-utt erance)가 빈번하게 사용되는데, 이는 음향 정보량이 제한적이어서 코덱 압축에 의한 왜곡에 더욱 민감할 수 있다[2]. 따라서 학습 데이터와 실제 서비스 환경 간의 코덱 특성 차이로 인한 성능 저하 문제를 규명하는 것은 기술의 실용화 관점에서 필수적인 과제이다.

본 논문에서는 2초 내외의 짧은 발화 조건에서 AASIST-L 모델의 코덱 및 채널 강건성을 정량적으로 분석한다. 공개 데이터셋인 Fake-or -Real(FoR)[3]을 활용하여 (1) 클린 환경, (2) 코덱 시뮬레이션 환경, (3) 재녹음 기반 현실 채널 환경에서의 성능 변화를 비교한다. 이를 통해 코덱 종류 및 비트레이트가 탐지 성능에 미치는 영향을 정량적으로 확인하고, 향후 실환경 강건성 개선을 위한 방향성을 제시하고자 한다.

Ⅱ. 데이터셋

본 논문에서는 공개 음성 딥페이크 탐지 데이터셋인 Fake-or-Real (FoR) 데이터셋을 사용한다. FoR은 실제 음성과 다양한 음성 합성 기법으로 생성된 합성 음성을 포함하며, clean 환경과 재녹음 기반 채널 환경을 함께 제공하여 탐지 모델의 강건성 평가에 활용된다. 본 논문은 FoR의 for-2sec와 for-rerec 버전을 사용한다.

for-2sec는 모든 음원을 2초 길이로 절단한 짧은 발화 데이터셋이다. 본 논문에서는 clean 조건의 for-2sec만을 사용하여 AASIST-L 모델을 학습하였다. 이는 실제 탐지 시스템이 주로 정제된 환경에서 학습되고, 이후 실제 서비스 환경에 배포되는 상황을 모사하기 위함이다.

for-rerec는 원본 음원을 스피커로 재생한 뒤 마이크로 다시 녹음하여 생성된 재녹음 기반 데이터셋으로, 실내 음향 특성, 하드웨어 왜곡, 배경 잡음 등 실제 통신 환경에서 발생 가능한 다양한 채널 변형이 포함되어 있다. 본 연구에서는 for-rerec test 세트를 사용해서 현실 채널 조건에서의 탐지 성능을 평가하였다.

평가 시나리오는 (1) for-2sec clean test, (2) for-2sec에 대해 FFmpeg 기반 AAC, MP3, Opus 트랜스코딩을 적용한 codec-sim test, (3) for-rerec test의 세가지로 구성된다. 이를 통해 clean 환경에서 학습된 모델이 코덱 압축 및 재녹음 기반 채널 변형에 노출될 때 발생하는 성능 변화를 정량적으로 비교한다.

Ⅲ. 실험 및 결과

본 연구는 짧은 발화(2초) 딥페이크 음성 탐지에서 코덱 압축 및 채널 변형에 따른 강건성(robustness)을 분석하기 위해 AASIST-L을 기반 탐지 모델로 사용한다. AASIST는 스펙트럼-시간 특징을 그래프 구조로 통합하고 그래프 어텐션을 통해 스푸핑 단서를 학습하는 anti-spoofing 모델로 제안되었다.[1] 또한 실제 서비스 환경에서는 전송/압축(코덱) 및 채널 조건 변화가 탐지 성능의 일반화 성능에 영향을 줄 수 있으며, 이러한 현실 변형을 고려한 평가 필요성이 반복적으로 제기되어 왔다.[2]

훈련 데이터는 FoR(Fake-or-Real) 데이터셋의 for-2sec(2초 절단) 버전을 사용하여 clean 조건에서 학습한다. 평가 지표는 EER(Equal Error Rate)을 사용하며, 이는 false acceptance rate와 false rejection rate가 동일해지는 지점의 오류율이다.[4]

본 연구의 평가 시나리오는 (i) 클린 환경, (ii) 코덱 시뮬레이션 환경(FFmpeg 기반 AAC/MP3/Opus 인코드-디코드), (iii) 재녹음 기반 현실 채

널 환경(for-rerec)으로 구성된다. 모든 실험에서 학습 조건은 for-2sec(clean)으로 고정하고, 평가 조건만 변경하여 도메인 시프트에 따른 성능 변화를 비교하였다. 또한 시나리오 2에서 코덱별 성능 저하를 정량화하기 위해 clean 기준 대비 EER 증가량을 Robustness Drop으로 정의한다. 학습은 AASIST-L 기본 설정을 따르되, 입력은 2초 음성을 단일 채널로 변환하고 16 kHz로 리샘플링하여 N = 32000샘플로 고정하였다. 최적화는 Adam(learning rate 0.0001, batch size 24)으로 수행하였으며, 개발(dev) EER이 개선되지 않을 경우 early stopping(patience=10)을 적용하였다. 모든 실험은 동일한 데이터 분할(train/dev/test)과 동일한 랜덤 시드에서 수행하여 조건 간 비교의 공정성을 유지하였다.

$$\Delta EER(cond) = EER(cond) - EER(clean)$$

시나리오 2의 코덱 시뮬레이션은 for-2sec test 음원을 대상으로 FFmpeg를 이용해 (i) 특정 코덱으로 인코딩한 뒤 (ii) 다시 PCM(WAV)으로 디코딩하는 트랜스코딩으로 생성하였다. 본 연구에서는 AAC/MP3는 64 kbps, Opus는 16/32 kbps 조건을 사용하였으며, 디코딩 후에는 모든 파일을 단일 채널 및 동일 샘플링 레이트로 정규화하여 모델 입력 형식을 일관되게 유지하였다.

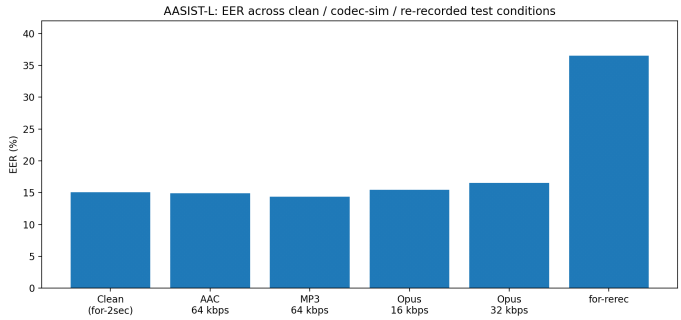

그림 1. 클린(for-2sec), 코덱 시뮬레이션(AAC/MP3/Opus), 재녹음(for-rerec) 조건에서 평가한 EER(%) 결과.

Test condition	EER(%)	ΔEER(%p)
Clean (for-2sec)	15.074	0.000
AAC 64k	14.890	-0.184
MP3 64k	14.338	-0.736
Opus 16k	15.441	+0.367
Opus 32k	16.544	+1.470
for-rerec	36.520	+21.446

표 1. AASIST-L 결과(학습: for-2sec 클린). 테스트 조건별 EER(%) 및 클린 대비 성능 변화량(ΔEER, %p).

표 1 및 그림 1은 AASIST-L의 조건별 EER을 요약한다. Clean 조건에서 EER은 15.074%로 측정되었다. 코덱 시뮬레이션 조건에서는 AAC 64 kbps(14.890%), MP3 64 kbps(14.338%), Opus 16 kbps(15.441%), Opus 32 kbps(16.544%)로 관측되었으며, 특히 Opus 32 kbps에서 clean 대비 성능 저하(+1.470%p)가 확인되었다. AAC/MP3 조건에서는 clean과 유사하거나 소폭 낮은 EER이 관측되었다. 다만 해당 차이는 1%p 미만으로 작아, 본 연구에서는 코덱 변환이 성능을 향상 시킨다고 단정하기보다는 평가 변동 범위 내의 변화로 해석하며, 향후 시드 반복 및 신뢰구간(또는 부트스트랩) 기반의 통계적 검증이 필요하다. 마지막으로 for-rerec 조건에서는 EER이 36.520%로 크게 증가(+21.446%p)하여, 단순 코덱 압축보다 "채널/재녹음"에 가까운 왜곡이 짧은 발화 탐지 성능에 훨씬 치명적일 수 있음을 시사한다. 이러한 경향은 실제 통신 환경에서 압축/전송/채널 변화가 탐지 성능을 저하시킨다는 기존 연구에서 보고된 경향과 유사한 방향성을 보인다.[2]

Ⅳ. 결론

본 연구는 2초 짧은 발화 딥페이크 음성 탐지에서 코덱 압축(Codec-sim)과 재녹음 기반 채널 변형(for-rerec)이 성능에 미치는 영향을 AASIST-L로 평가하였다. 학습은 for-2sec(clean)으로 고정하고, 테스트 조건만 변경하여 도메인 시프트에 따른 강건성 변화를 비교했다. 결과적으로 Clean EER은 15.074%였으며, 코덱 시뮬레이션에서는 코덱/설정에 따라 영향이 달랐다. 특히 Opus 32 kbps에서 EER 16.544% (+1.470%p)로 가장 큰 성능 저하가 관측되어, "압축은 항상 동일하게 성능을 떨어뜨린다"가 아니라 코덱 민감도는 코덱 종류와 비트레이트에 따라 달라질 수 있음을 확인했다. 반면 AAC/MP3는 clean과 유사한 수준을 보여, 단일 코덱 조건만으로 강건성을 일반화하기 어렵다는 점을 시사한다. 가장 큰 변화는 for-rerec에서 EER 36.520%(+21.446%p)로 급증한 것으로, 단순 코덱 압축보다 채널/재녹음에 가까운 현실 변형이 짧은 발화 탐지 성능에 훨씬 치명적일 수 있음을 보여준다. 향후에는 코덱/채널 변형을 학습 단계에 포함하거나, 전송 통계(지터·손실 등)와 음향 특징을 결합하는 방식으로 실환경 강건성을 개선하는 실험을 진행할 계획이다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원을 받아 수행된 연구임 (IITP-2026-RS-2024-00438430).

참 고 문 헌

[1] Jung, Jee-weon, et al. "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks." ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2022.

[2] Liu, Xuechen, et al. "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild." IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2023): 2507-2522.

[3] R. Reimao and V. Tzerpos, "For: A Dataset for Synthetic Speech Detection," in Proc. SpeD, 2019, pp. 1-10.

[4] Delgado, Héctor, et al. "ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan." arXiv preprint arXiv:2109.00535 (2021).