가짜 음성 합성 탐지 기술 개발

프로젝트 수행 기간: 2024.05~2024.08

최우수상 (정보통신기획평가원장상), SW중심대학 디지털 경진대회 AI부분

데이터 분석

학습 및 평가 데이터 분석

- 훈련 데이터: 55438개의 32kHz로 샘플링 된 오디오(.ogg) 샘플
- 방음 환경에서 녹음된 Real 오디오 샘플: 27620개
- 방음 환경을 가정한 Fake 오디오 샘플: 27818개
- Real/Fake의 샘플 비율은 balance하지만, 각각 샘플 오디오 길이는 상이함

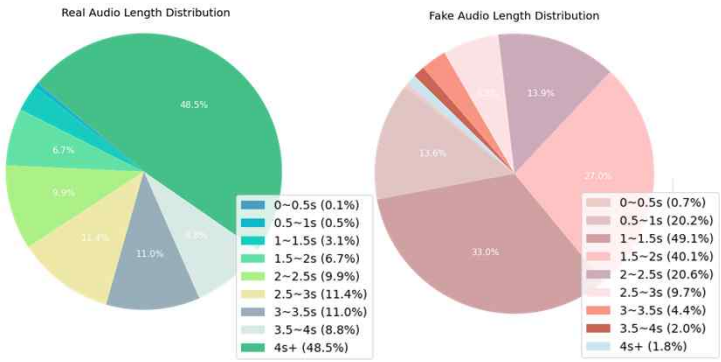


그림 1. 훈련 데이터 오디오 샘플 길이 비교

- 평가 데이터 : 50000개의 5초 분량의 32kHz로 샘플링 된 오디오(.ogg) 샘플
- 방음이 아닌 환경도 존재, 음성에 다양한 잡음이 더해짐
- 화자가 1명 혹은 2명일 수 있으며 2명 모두 Real이거나 2명 모두 Fake일 수 있음

그림 2. 평가 데이터 분석

데이터셋 구축

노이즈 데이터 구축

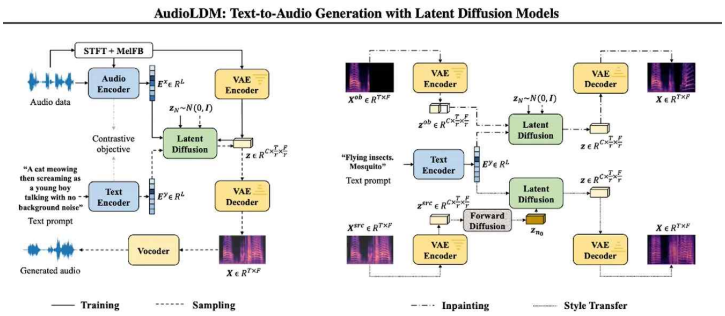


그림 3. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models

- Text를 Audio로 변환하는 잠재적 확산 모델(LDM)
- CLAP(latent continuous audio representations)에서 연속적 오디오 표현
- 텍스트 프롬프트를 입력으로 받아 Condition으로 적용하여 오디오를 예측함

데이터셋 구축

Table1. Dataset 설계

Label	화자 수	Real	Fake	Real Scale	Fake Scale	Train Noise Scale	Validation Noise Scale
[1,1]	2명	1명	1명	1명	1명	0.1~0.5	0.3~0.5
[1,0]	2명	X	2명	X	2명	0.1~0.5	0.3~0.5
	1명	X	1명	X	1명	0.1~0.5	0.3~0.5
[0,1]	2명	2명	X	2명	X	0.1~0.5	0.3~0.5
	1명	1명	X	1명	X	0.1~0.5	0.3~0.5
[0 ,0]	X	X	X	X	X	0.1~0.5	0.3~0.5

- Label [0,0] 인 경우 생성 모델로 생성한 잡음만을 부여한다.
- 이외의 경우 Real과 Fake의 음성 스케일을 0.1~0.5 범위 내에서 지정하며 Noise의 스케일의 범위도 동일하게 부여한다.

모델 알고리즘 및 손실함수

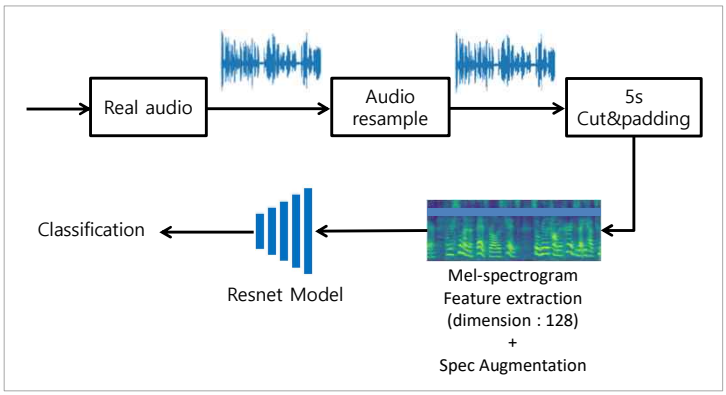
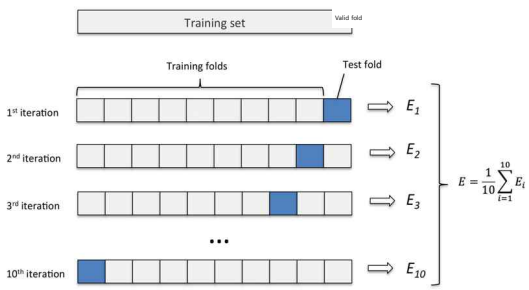


그림 4. 학습, 검증 데이터 및 노이즈에 대한 전처리

- 모든 오디오에 대해 16kHz로 샘플링, 5초 분량의 오디오 길이로 통일
- Mel-Spectrogram으로 128차원의 주파수 특징 추출 후 Resnet 모델 학습

K-fold cross validation & Ensemble

그림5. k-fold 교차 검증과 앙상블 기법



- 음성 데이터셋을 특정 범위에 따라 10개의 validation 폴드로 나눈 후, 폴드 별 교차 검증
- 폴드 별로 검증 데이터셋에 따라 모델 평가, 성능 측정치 평균 후 최종 모델 성능 추정

결론

- 모델의 크기를 키우거나 더 많은 feature를 추출하는 것보다, 데이터 증강을 통해 더 효율적이고 실질적인 성능 개선을 확인
- 노이즈가 있는 환경에서의 모델 성능은 학습 데이터에 적절한 노이즈를 추가하여 모델이 다양한 환경에 적응할 수 있도록 설계하는 것이 중요한 역할을 한 것이라 판단