2024 SW중심대학 디지털 경진대회

AI부문

조선대학교

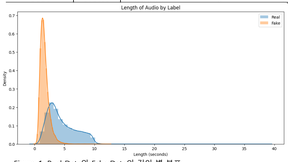
어떻게 너의 목소리를 잊겠어(백찬형, 문지환)

1. 데이터 분석

1. **Train Data(Fake and Real)**
   - Train Data는 32kHz로 샘플링된 55438개의 샘플로 구성됨
   - Noise가 없이 깨끗한 음성만 존재
2. **Unlabel Data**
   - Unlabel Data 또한 32kHz로 샘플링 되었으며, 1264개의 샘플로 구성됨
   - Test Data와 동일하게 Noise가 있으며, 0~2명의 발화자로 구성된 5초 길이의 음성 데이터

표 1. 데이터셋 구성 및 세부 정보

Data	Label	Sample	Length(s)	Environment	kHz	Speakers
Train	Fake	27818	~8	clean	32	1
	Real	27620	~35	clean	32	1
Unlabel	None	1264	5	noisy	32	0~2
Test	None	50000	5	Noisy	32	0~2

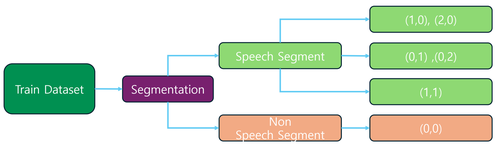


- 5초 이상의 진짜 음성 파일 9807개
- 5초 이하의 진짜 음성 파일 17813개
- 5초 이상의 가짜 음성 파일 74개
- 5초 이하의 가짜 음성 파일 27744개

Figure 1. Real Data와 Fake Data의 길이 별 분포

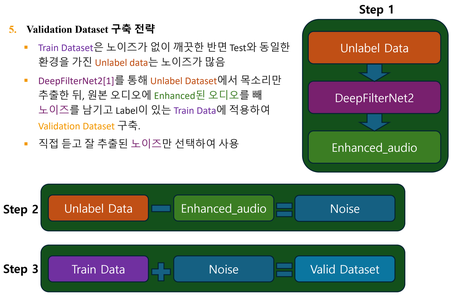
2. Dataset 구축

4. **어떻게 Unbalance를 해결?**
   - 모든 Train data를 Segmentation을 통해 발화 부분만 추출
   - 5초 단위로 자르고, 0.42초 미만이면 버림
   - Test Dataset도 Segmentation을 수행 후 발화 부분이 검출되지 않으면 (0,0)으로 제출
   - 5초 미만 음성은 5초로 padding된 음성에 랜덤한 위치에서 실행
   - 이 과정을 거친 Train Dataset이 (1,0), (0,1) ,(1,1),(2,0),(0,2) label을 가지도록 구축
   - (0,0)은 Train dataset에서 Segment가 검출되지 않은 영역을 골라 (0,0)으로 구축



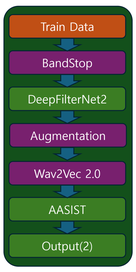
2. Dataset 구축

5. **Validation Dataset 구축 전략**
   - Train Dataset은 노이즈가 없이 깨끗한 반면 Test와 동일한 환경을 가진 Unlabel data는 노이즈가 많음
   - DeepFilterNet2[1]를 통해 Unlabel Dataset에서 목소리만 추출한 뒤, 원본 오디오에 Enhanced된 오디오를 빼 노이즈를 남기고 Label이 있는 Train Data에 적용하여 Validation Dataset 구축.
   - 직접 듣고 잘 추출된 노이즈만 선택하여 사용



3. 학습 전략

1. **어떻게 노이즈에 강인한 모델을 만들까?**
   - 1. Augmentation으로 노이즈를 최대한 생성하는 것
   - 2. Speech Enhance model을 통해 노이즈를 없애는 것
   - 크게 두 가지 방법론에서 2번을 선택
2. **Speech Enhance Model(DeepFilterNet2)**
   - DeepFilterNet2는 DNS 4 challenge Dataset을 가지고 학습된 Pretrain model을 사용
   - Speech Enhance model을 통해 Task를 노이즈가 적은 ASVSpoof2021 Task와 유사하게 도메인 적응을 시도
   - 이에 따라 ASVSpoof2021에서 좋은 성능을 보인 Wav2Vec2.0 + AASIST 도입



3. 학습 전략

3. **Speech Enhance Model의 단점**
   - Speech Enhance Model도 생성 모델에 가깝기 때문에 노이즈가 심할 때 Real Data의 일부 구간이 Fake처럼 들리는 Artifact 생김
4. **Speech Enhance Model의 단점을 Focus!**
   - 일부 구간이 Fake처럼 들리는 Artifact가 생긴다면 그 부분을 전처리를 통해서 Train에 구현하자
   - 1. GainTransition : 급격히 소리가 줄어들거나 소리가 커지는 Aug
   - 2. RandomPitchShift : 자체 구현(랜덤적인 위치에서 Pitch를 극하게 내리거나 올림, 1~3초 사이의 랜덤한 시간만 내린 후 원상 복귀)
   - 3. BandStop : BandStop을 거친 뒤 DeepFilterNet2를 거치면 Artifact가 생성되는 측면을 관찰할 수 있음

4. Model

1. **Wav2Vec 2.0 + AASIST**
   - Enhance Model을 활용하여 Noise를 없앴을 때 ASVSpoof 2021과 비슷하게 접근해도 되겠다는 판단
   - 따라서 Hugging Face의 Wav2Vec 2.0 Feature Extractor와 AASIST를 선택

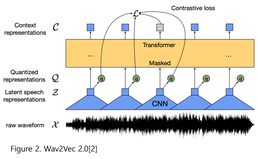
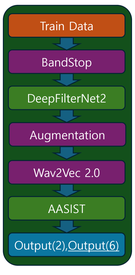


Figure 2. Wav2Vec 2.0[2]

2. **기존 AASIST**
   - 기존 AASIST는 softmax를 통해 2개의 Output을 최종적으로 도출
3. **수정된 AASIST**
   - 기존엔 화자가 2명이여도 1명과 동일한 (1,0),(0,1)로 뱉어야함
   - Dataloader를 {batch,2}, {batch,6}으로 별도록 설계
   - AASIST의 output을 {batch,2},{batch,6}으로 나오도록 수정
   - Fake와 Real의 0~1 확률과 6가지 label을 동시에 맞히도록 유도
     => 모델이 화자 명수를 반영하여 예측하기 위함
   - Loss: MultiLabelSoftmargin x 0.65(2 label) + crossEntropy x 0.35(6label)



전체적인 학습 메커니즘

5. Model 및 실험 결과

1. **실험 결과**
   1) Audio Segmentation을 통한 Train Dataset 구축
   2) Augmentation(gain,pitchshift, bandstop) 추가
   3) 6label 구조
   4) Temperature scaling
   - 4가지가 score 상승에 큰 기여.

표 2. 모델 성능 및 개선 사항 적용 실험 결과

Model	Public Score	적용 사항
Wav2Vec2	0.29081	(2,0),(0,2),(1,1) 구축
	0.24942	Audio segmentation(Train),(0,0)은 padding+noise
Wav2Vec2+AASIST	0.20451	Augmentation(gain,pitchshift,bandstop) 추가
Wav2Vec2+AASIST (모델구조 수정)	0.18683	6label 구조
	0.15557	Temperature scaling 추가

6. 결론

1. **결론**
   - 노이즈를 생성하거나, 노이즈를 없애거나 그 둘 중 노이즈를 없애는 것으로 결정
   - 노이즈를 없애는 과정에서 생긴 artifact를 개선하기 위해 전처리 기법 사용
   - 노이즈가 없다는 가정 하에 SOTA 모델인 wav2vec 2.0+ aasist 도입
   - 모델의 calibration 성능을 높이기 위한 Temperature scaling 추가
   - 앙상블을 하지 않고도 Public score 기준 약 0.15557, Private score 기준 약 0.15891의 성능을 보임
2. **향후 계획**
   - 노이즈에 강인한 Anti-spoofing 방법론을 위한 Speech Enhancement Model 개선
   - Speech Enhancement Model로 발생된 Artifact를 보완하기 위한 전처리 연구
3. **참가 소감**
   - 마지막으로 참가하는 대회인만큼 열심히 준비해 보람찬 대회였습니다.

팀장(백찬형) 번호 : 010-2856-0288
팀원(문지환) 번호 : 010-7152-1376

참조[1]: S. M. Lauritsen et al., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations", *arXiv 2006.11477v3 [cs.CL] 22 Oct 2020*
참조[2]: S. M. Lauritsen et al., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations", *arXiv 2006.11477v3 [cs.CL] 22 Oct 2020*
참조[3]: J. weon Jung et al., "AASIST: AUDIO ANTI-SPOOFING USING INTEGRATED SPECTRO-TEMPORAL GRAPH ATTENTION NETWORKS", arXiv-2110.01200v1 [eess.AS] 4 Oct 2021

주최 | 과학기술정보통신부 주관 | IITP 정보통신기획평가원 SW 중심대학협의회