UniSpeech-SAT: Universal Speech Representation Learning with Speaker-Aware Pre-Training

https://arxiv.org/pdf/2110.05752

0. Introduction

- 기존 음성 self-supervised 모델(Wav2Vec 2.0, HuBERT 등) → 발화 내용 중심 학습
- 화자 정보(speaker identity) 손실 문제 존재
- UniSpeech-SAT → Speaker-Aware Pre-Training 도입
- Content + Speaker representation 동시 학습 목표
- 주요 기여
  - Speaker recognition objective 통합
  - Dual objective 설계로 균형 학습
  - 다양한 downstream task에서 범용 성능 확보

1. Overview

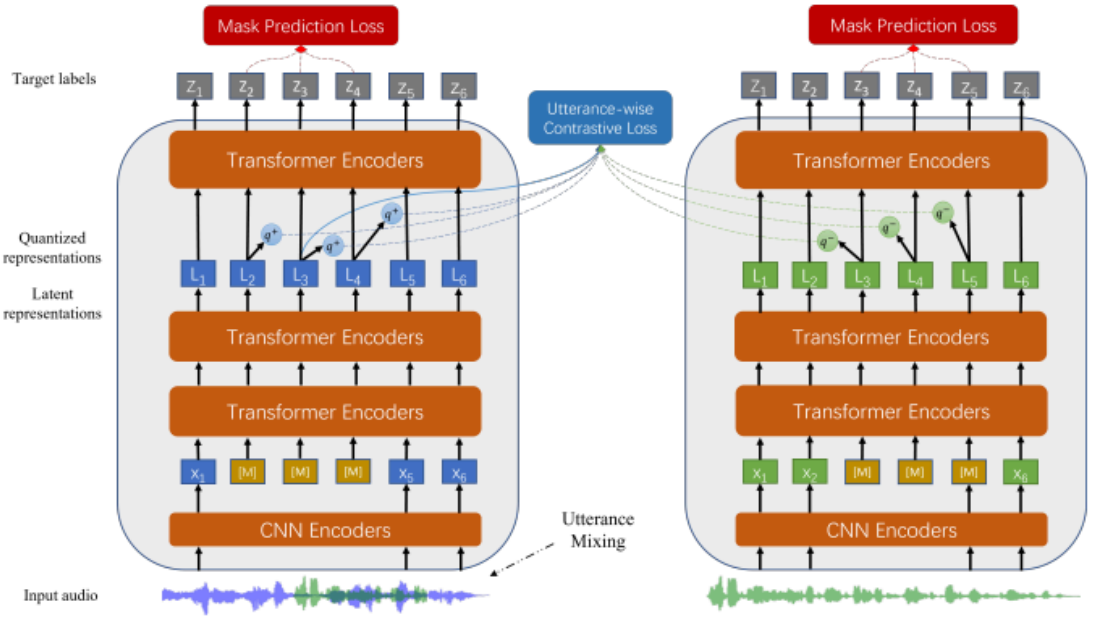
- UniSpeech의 확장 버전
- Transformer encoder 기반 self-supervised 구조
- Masked feature prediction + Speaker embedding branch 구성
- Content / Speaker 정보 동시 학습

- labeled + unlabeled 데이터 모두 사용 가능
- ASR, Speaker Verification, Emotion Recognition 등 다중 태스크 적용
- 목표: 하나의 pre-trained 모델로 범용 음성 표현 학습

2. Challenges

- 기존 모델 → content 중심 → speaker 정보 손실
- Speaker task 성능 저하 발생
- Content vs Speaker 정보 간 trade-off 존재
- Speaker label 없는 unlabeled 데이터 학습 어려움
- Multi-task 학습 시 gradient 간섭(interference) 문제
- 안정적 joint optimization 필요

3. Method



- Dual objective pre-training 구조
  - LcontentL_{content}Lcontent: masked prediction

    - LspeakerL_{speaker}Lspeaker: speaker classification loss
- 입력 음성 → feature extractor → 일부 마스킹 → Transformer encoder 통과
- Speaker embedding branch 추가 → speaker identity 학습
- Adversarial training + normalization → content/speaker 간 간섭 최소화
- 최종 손실:

  Ltotal=Lcontent+λ·LspeakerL_{total} = L_{content} + \lambda \cdot L_{speaker}Ltotal=Lcontent+λ·Lspeaker
- λ로 두 목표 균형 조정
- labeled/unlabeled 음성 모두 활용 가능
- fine-tuning 시 downstream task 맞춤 조정

4. Experiments

- 데이터셋: LibriSpeech 960h, VoxCeleb 1/2, CN-Celeb, Switchboard 등
- 비교모델: Wav2Vec 2.0, HuBERT, UniSpeech
- 태스크: ASR / Speaker Verification / Emotion Recognition / Speech Translation
- 지표: WER(ASR), EER(Speaker Verification)
- fine-tuning 비율: 10%, 100%
- Ablation: λ 변화 영향, speaker objective 제거 시 효과 분석

5. Results

Method	#Params	Corpus	Speaker			Content					Semantics			ParaL.	Overall
			SID	ASV	SD	PR	ASR (WER)		KS	QbE	IC	SF		ER	
			Acc ↑	EER ↓	DER ↓	PER ↓	w/o ↓	w/ LM ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Score ↑
FBANK	-	-	8.5E-4	9.56	10.05	82.01	23.18	15.21	8.63	0.0058	9.10	69.64	52.94	35.39	44.2
PASE+ [14]	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	16.62	82.54	0.0072	29.82	62.14	60.17	57.86	57.5
APC [8]	4.11M	LS 360 hr	60.42	8.56	10.53	41.98	21.28	14.74	91.01	0.0310	74.69	70.46	50.89	59.33	67.6
VQ-APC [10]	4.63M	LS 360 hr	60.15	8.72	10.45	41.08	21.20	15.21	91.11	0.0251	74.48	68.53	52.91	59.66	67.2
NPC [11]	19.38M	LS 360 hr	55.92	9.40	9.34	20.20	13.91	43.81	88.96	0.0246	69.44	72.79	48.44	59.08	67.0
Mockingjay [12]	85.12M	LS 360 hr	32.29	11.66	10.54	22.82	15.48	70.19	83.67	6.6E-04	34.33	61.59	58.89	50.28	56.1
TERA [13]	21.33M	LS 360 hr	57.57	15.89	9.96	18.17	12.16	49.17	89.48	0.0013	58.42	67.50	54.17	56.27	64.2
modified CPC [2]	1.84M	LL 60k hr	39.63	12.86	10.38	42.54	20.18	13.53	91.88	0.0326	64.09	71.19	49.91	60.96	65.1
wav2vec [3]	32.54M	LS 960 hr	56.56	7.99	9.90	31.58	15.86	11.00	95.59	0.0485	84.92	76.37	43.71	59.79	71.5
vq-wav2vec [4]	34.15M	LS 960 hr	38.80	10.38	9.93	33.48	17.71	12.80	93.38	0.0410	85.68	77.68	41.54	58.24	69.3
wav2vec 2.0 Base [5]	95.04M	LS 960 hr	75.18	5.74	6.02	6.08	6.43	4.79	96.23	0.0233	92.35	88.30	24.77	63.43	80.3
HuBERT Base [6]	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	**6.42**	**4.79**	96.30	0.0736	98.34	88.53	25.20	64.92	82.0
UniSpeech-SAT Base	94.68M	LS 960 hr	85.76	**4.31**	4.41	5.40	6.75	4.86	96.75	0.0927	98.58	88.98	23.56	66.04	83.0
– contrastive loss	94.68M	LS 960 hr	84.74	4.61	4.72	5.22	6.80	5.17	96.79	0.0956	98.31	88.56	24.00	65.60	82.8
– utterance mixing	94.68M	LS 960 hr	85.97	4.35	5.87	5.06	7.04	5.05	96.88	0.0866	98.10	88.50	24.52	65.97	82.7
UniSpeech-SAT Base+	94.68M	CD 94k hr	**87.59**	4.36	**3.80**	**4.44**	6.44	4.88	**97.40**	**0.1125**	**98.84**	**89.76**	**21.75**	**68.48**	**84.0**
wav2vec 2.0 Large [5]	317.38M	LL 60k hr	86.14	5.65	5.62	4.75	3.75	3.10	96.6	0.0489	95.28	87.11	27.31	65.64	82.1
HuBERT Large [6]	316.61M	LL 60k hr	90.33	5.98	5.75	3.53	**3.62**	**2.94**	95.29	0.0353	98.76	89.81	21.76	67.62	83.5
UniSpeech-SAT Large	316.61M	CD 94k hr	**95.16**	**3.84**	**3.85**	**3.38**	3.99	3.19	**97.89**	**0.0836**	**99.34**	**92.13**	**18.01**	**70.68**	**85.6**

- ASR: UniSpeech 대비 WER 5~10% 향상
- Speaker Verification: EER 20% 감소
- 단일 pre-trained 모델 → 다중 태스크 전이 가능
- Speaker objective 제거 시 speaker task 성능 급감
- λ 값 조정 중요 → 과도 시 content 성능 하락
- Speaker-aware 설계 → representation 범용성 강화 확인

6. Insight

- Content + Speaker 동시 학습 가능성 입증
- 범용 음성 표현(universal representation) 학습 진전
- Speaker 정보 손실 문제 구조적으로 해결
- 후속 연구 방향
  - Emotion 등 추가 화자 특성 확장
  - Cross-lingual 강건성 향상
  - Text-supervised 하이브리드 구조 탐색
- UniSpeech-SAT → speaker-aware 확장형 self-supervised 음성 모델로 평가