HuBERT : Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units

https://arxiv.org/abs/2106.07447

0. Introduction

- 이 논문은 음성 데이터를 위한 self-supervised representation learning 문제를 다룸
- 기존 음성 표현 학습 방식의 문제점으로, (1) 입력 발화에 여러 sound unit 이 포함됨, (2) pre-training 단계에서는 사전 정의된 음성 단위가 없음, (3) sound unit 들이 가변 길이이고 명시적 segmentation이 어렵다는 점을 지적
- 이러한 어려움을 해결하기 위해 “마스킹된 구간의 hidden unit 예측”이라는 아이디어를 도입하고, offline clustering을 통해 target label을 만들고 masked-prediction loss로 학습하는 방식을 제안
- 이 논문의 핵심 기여는 명시적 레이블 없이도 음성의 잠재 표현을 효과적으로 학습할 수 있는 실용적인 self-supervised 방식 제안

1. Overview

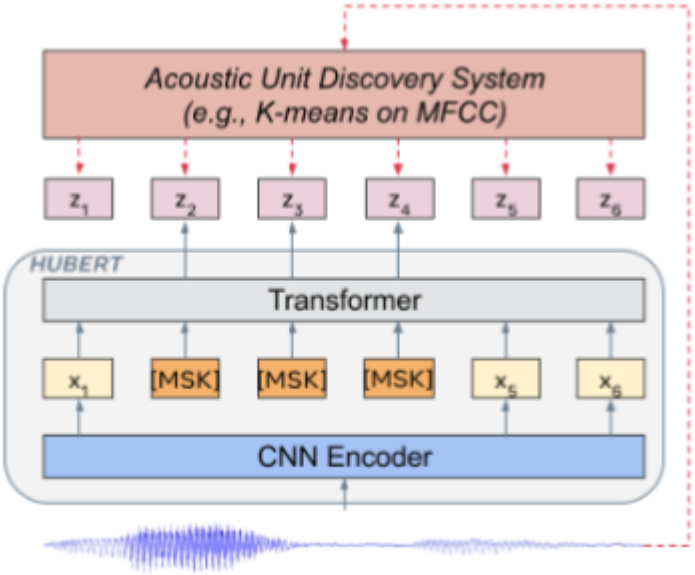
- 제안 방식은 두 단계로 구성
  - offline clustering을 통해 각 프레임 또는 segment에 cluster ID를 할당하여 pseudo label 생성
  - 입력 음성의 일부 구간을 마스킹하고, 해당 구간의 cluster ID를 예측하도록 학습
- 이 구조는 텍스트용 사전학습 모델 BERT와 유사한 masked prediction 구조를 가짐
- 학습된 speech representation은 음향 정보와 언어 정보를 동시에 포착하도록 설계됨

- LibriSpeech 960 h 및 Libri-light 60,000 h 규모의 대규모 데이터셋을 활용해 실험 수행
- 사전 레이블 없이도 다양한 음성 처리 downstream task에 전이 가능하다는 점이 핵심 장점

2. Challenges

- 음성은 텍스트와 달리 명확한 고정 단위가 없어 segmentation 자체가 어려움
- unsupervised clustering 기반 pseudo label 은 noise가 많고 실제 언어 단위와 정확히 대응되지 않을 수 있음
- masked-prediction 방식이 잡음 환경이나 다양한 화자 조건에서도 안정적으로 작동할 수 있을지는 실험 조건에 따라 달라질 수 있음
- downstream task 전이 시 representation의 일반화 성능이 데이터 분포에 영향을 받을 가능성 존재

3. Method



- Offline clustering 단계에서 원시 음성을 프레임 또는 작은 단위로 분할한 뒤 k-means clustering 수행
- 각 프레임 또는 segment에 대해 cluster ID를 target label로 사용

- 입력 음성의 일부 구간을 무작위로 마스킹하고, masked 구간의 cluster ID를 예측하도록 학습
- Loss는 masked region에 대해서만 계산
- Clustering 단계와 representation 학습 단계를 반복 수행하며 점진적으로 성능을 개선
- 모델 구조는 BERT 유사 구조이며 연속적인 음향 feature를 입력으로 사용

4. Experiments

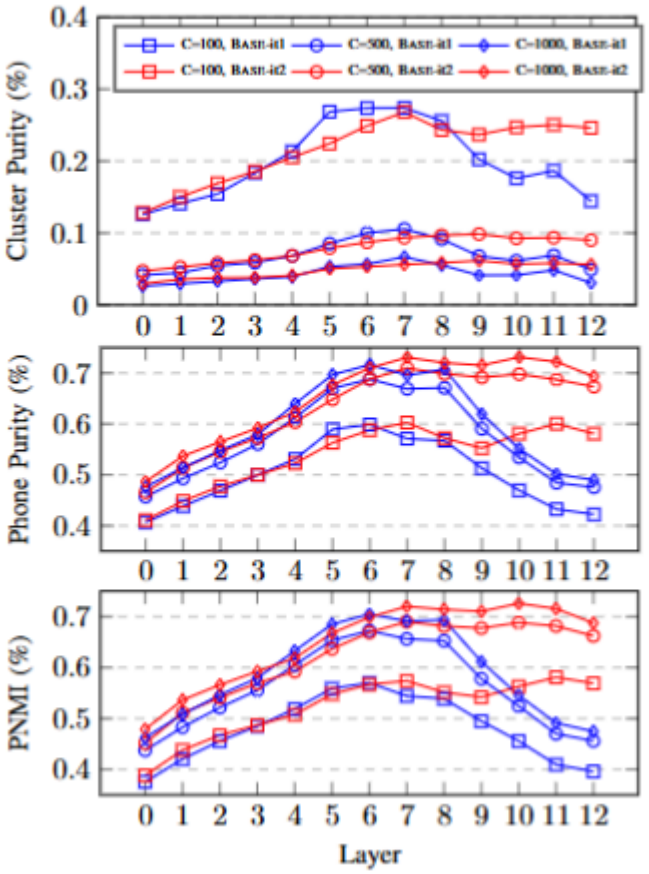
- 사용 데이터셋은 LibriSpeech 960 h 와 Libri-light 60,000 h
- Fine-tuning 데이터 크기를 10 min, 1 h, 10 h, 100 h, 960 h 로 다양하게 설정
- Baseline은 wav2vec 2.0 등 당시 최상위 self-supervised 음성 모델
- 평가 지표는 WER 사용
- dev-other, test-other 와 같은 난이도 높은 세트를 중심으로 성능 비교

5. Results

		BASE	LARGE	X-LARGE
CNN Encoder	strides	5, 2, 2, 2, 2, 2, 2		
	kernel width	10, 3, 3, 3, 3, 2, 2		
	channel	512		
Transformer	layer	12	24	48
	embedding dim.	768	1024	1280
	inner FFN dim.	3072	4096	5120
	layerdrop prob	0.05	0	0
	attention heads	8	16	16
Projection	dim.	256	768	1024
Num. of Params		95M	317M	964M

Model	Unlabeled Data	LM	dev-clean	dev-other	test-clean	test-other
		10-min labeled				
DiscreteBERT [51]	LS-960	4-gram	15.7	24.1	16.3	25.2
wav2vec 2.0 BASE [6]	LS-960	4-gram	8.9	15.7	9.1	15.6
wav2vec 2.0 LARGE [6]	LL-60k	4-gram	6.3	9.8	6.6	10.3
wav2vec 2.0 LARGE [6]	LL-60k	Transformer	4.6	7.9	4.8	8.2
HUBERT BASE	LS-960	4-gram	9.1	15.0	9.7	15.3
HUBERT LARGE	LL-60k	4-gram	6.1	9.4	6.6	10.1
HUBERT LARGE	LL-60k	Transformer	4.3	7.0	4.7	7.6
HUBERT X-LARGE	LL-60k	Transformer	4.4	6.1	4.6	6.8
		1-hour labeled				
DeCoAR 2.0 [50]	LS-960	4-gram	-	-	13.8	29.1
DiscreteBERT [51]	LS-960	4-gram	8.5	16.4	9.0	17.6
wav2vec 2.0 BASE [6]	LS-960	4-gram	5.0	10.8	5.5	11.3
wav2vec 2.0 LARGE [6]	LL-60k	Transformer	2.9	5.4	2.9	5.8
HUBERT BASE	LS-960	4-gram	5.6	10.9	6.1	11.3
HUBERT LARGE	LL-60k	Transformer	2.6	4.9	2.9	5.4
HUBERT X-LARGE	LL-60k	Transformer	2.6	4.2	2.8	4.8
		10-hour labeled				
SlimIPL [54]	LS-960	4-gram + Transformer	5.3	7.9	5.5	9.0
DeCoAR 2.0 [50]	LS-960	4-gram	-	-	5.4	13.3
DiscreteBERT [51]	LS-960	4-gram	5.3	13.2	5.9	14.1
wav2vec 2.0 BASE [6]	LS-960	4-gram	3.8	9.1	4.3	9.5
wav2vec 2.0 LARGE [6]	LL-60k	Transformer	2.4	4.8	2.6	4.9
HUBERT BASE	LS-960	4-gram	3.9	9.0	4.3	9.4
HUBERT LARGE	LL-60k	Transformer	2.2	4.3	2.4	4.6
HUBERT X-LARGE	LL-60k	Transformer	2.1	3.6	2.3	4.0
		100-hour labeled				
IPL [12]	LL-60k	4-gram + Transformer	3.19	6.14	3.72	7.11
SlimIPL [54]	LS-860	4-gram + Transformer	2.2	4.6	2.7	5.2
Noisy Student [61]	LS-860	LSTM	3.9	8.8	4.2	8.6
DeCoAR 2.0 [50]	LS-960	4-gram	-	-	5.0	12.1
DiscreteBERT [51]	LS-960	4-gram	4.0	10.9	4.5	12.1
wav2vec 2.0 BASE [6]	LS-960	4-gram	2.7	7.9	3.4	8.0
wav2vec 2.0 LARGE [6]	LL-60k	Transformer	1.9	4.0	2.0	4.0
HUBERT BASE	LS-960	4-gram	2.7	7.8	3.4	8.1
HUBERT LARGE	LL-60k	Transformer	1.8	3.7	2.1	3.9
HUBERT X-LARGE	LL-60k	Transformer	1.7	3.0	1.9	3.5

Model	Unlabeled Data	LM	dev-clean	dev-other	test-clean	test-other
		Superivsed				
Conformer L [62]	-	LSTM	-	-	1.9	3.9
		Self-Training				
IPL [12]	LL-60k	4-gram + Transformer	1.85	3.26	2.10	4.01
Noisy Student [61]	LV-60k	LSTM	1.6	3.4	1.7	3.4
		Pre-Training				
wav2vec 2.0 LARGE [6]	LL-60k	Transformer	1.6	3.0	1.8	3.3
pre-trained Conformer XXL [40]	LL-60k	LSTM	1.5	3.0	1.5	3.1
		Pre-Training + Self-Training				
wav2vec 2.0 + self-training [63]	LL-60k	Transformer	1.1	2.7	1.5	3.1
pre-trained Conformer XXL + Noisy Student [40]	LL-60k	LSTM	1.3	2.6	1.4	2.6
		This work (Pre-Training)				
HUBERT LARGE	LL-60k	Transformer	1.5	3.0	1.9	3.3
HUBERT X-LARGE	LL-60k	Transformer	1.5	2.5	1.8	2.9



teacher	C	PNMI	dev-other WER (%) $\alpha = 1.0$	$\alpha = 0.5$	$\alpha = 0.0$
Chenone (supervised top-line)	8976	0.809	10.38	9.16	9.79
K-means on MFCC	50	0.227	18.68	31.07	94.60
	100	0.243	17.86	29.57	96.37
	500	0.276	18.40	33.42	97.66
K-means on BASE-it1-layer6	500	0.637	11.91	13.47	23.29
K-means on BASE-it2-layer9	500	0.704	10.75	11.59	13.79

TABLE V: The effect of the training objective and clustering quality on performance. C refers to the number of units, and α is the weight for masked frames.

teacher	WER
K-means {50,100}	17.81
K-means {50,100,500}	17.56
Product K-means-0-100	19.26
Product K-means-1-100	17.64
Product K-means-2-100	18.46
Product K-means-{0,1,2}-100	16.73

TABLE VI: Cluster ensembles with k-means and product k-means.

- HuBERT는 wav2vec 2.0 과 유사하거나 더 나은 성능을 달성
- 대규모 모델 기준으로 dev-other 및 test-other 에서 상대 WER 감소 달성
- fine-tuning 데이터가 적은 환경에서도 안정적인 성능 향상 확인
- 다양한 데이터 규모 조건에서 robust한 speech representation을 제공함을 실험적으로 검증

6. Insight

- 명시적 레이블 없이도 고품질 음성 표현 학습이 가능함을 입증했다는 점에서 실용적 의미가 큼
- offline clustering과 masked prediction을 결합한 구조는 음향 정보와 언어 정보를 동시에 학습하려는 효과적인 접근
- pseudo label 기반 학습 특성상 clustering 품질에 따라 성능이 민감하게 영향을 받을 수 있다는 한계 존재
- 레이블이 부족한 언어, 방언, 특정 도메인 음성 데이터에 특히 유용한 구조로 해석 가능
- 음성 인식, 음성 합성, 음성 분류 등 다양한 downstream task 확장 가능성 존재