AV-HuBERT: Self-Supervised Audio-Visual Speech Representation Learning

https://arxiv.org/abs/2201.02184

0. Introduction

- 음성 인식은 전통적으로 오디오 신호만을 기반으로 학습되어 왔음
- 실제 사람은 청각 정보뿐 아니라 시각적 단서(입술 움직임 등)를 함께 사용해 음성을 이해함
- 특히 잡음 환경(noisy environment)에서는 시각적 단서가 중요한 역할을 함
- 기존의 오디오 전용 자가 지도 학습 모델(HuBERT 등)은 오디오 표현 학습에서는 성과가 있었으나 시각적 정보 활용은 제한적이었음
- AV-HuBERT는 HuBERT 프레임워크를 확장하여 오디오 + 비디오(입술 움직임)를 함께 학습하는 모델임
- 핵심 아이디어 : 오디오와 비디오의 상호 보완적 특징을 활용하여 더 강력한 음성 표현을 자가 지도 방식으로 학습함
- 목표 :
    - 라벨이 부족한 상황에서도 성능 높은 음성 인식 달성
    - 멀티모달 입력을 활용하여 잡음 환경에서도 강건한(srobust) 모델 구현
    - 오디오 전용 학습 모델의 한계를 극복하고 시각·청각 통합 학습의 가능성을 제시

1. Overview

- AV-HuBERT는 기존 HuBERT(Hidden Unit BERT) 구조를 기반으로 오디오 전용에서 오디오-비디오 멀티모달 학습으로 확장한 모델임
- 입력으로 오디오 스펙트로그램과 비디오 프레임(입술 움직임 중심)을 동시에 받아서 처리함
- 모델 구조는 크게 세 단계 :

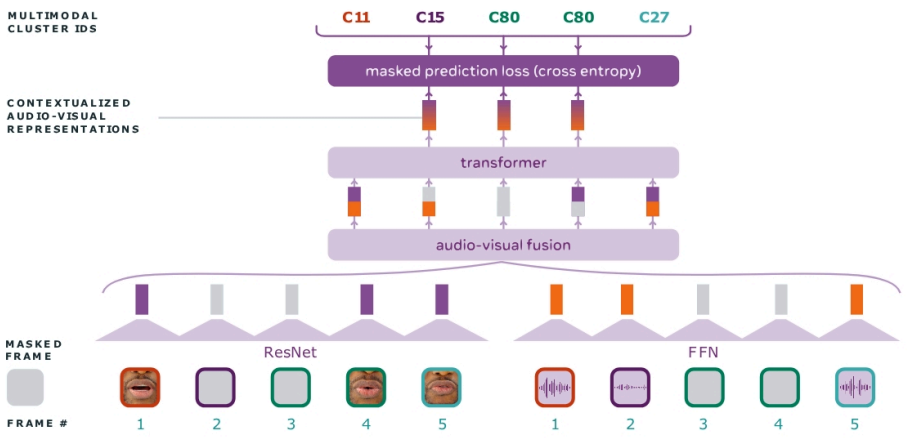
    - 멀티모달 인코더 : 오디오와 비디오 각각에 대해 특징을 추출하고 이를 융합(fusion)
    - 클러스터링 기반 타겟 생성 단계 : 라벨이 없는 데이터를 클러스터링해 pseudo-label(가짜 라벨)을 생성
    - 마스크 예측 단계 : 일부 입력을 마스킹하고, 그 부분을 맞추도록 학습
- 학습된 표현은 다운스트림 작업(ASR, lip reading, AVSR 등)에 활용 가능함
- 오디오와 비디오 모달리티가 상호 보완적으로 작용해 성능 향상
    - 오디오가 불완전할 때 영상이 보완
    - 영상 정보가 제한적일 때 오디오가 보완
- 기존 멀티모달 음성 모델 대비 차별점
    - 완전히 자가 지도 학습 기반으로 설계됨 → 라벨이 부족한 상황에서도 학습 가능
    - 단계적 학습 전략 사용 (Iterative refinement)
    - 단순한 early fusion이 아니라, representation level에서의 융합으로 더 일반화된 표현 학습 가능
- AV-HuBERT는 시각-청각 멀티모달 음성 표현 학습에서 새로운 SOTA 수준 성능을 달성함

2. Challenges

- 멀티모달 데이터 정렬(synchronization) 문제
    - 오디오와 영상은 시간 단위가 다름
    - 프레임 레이트 불일치, 발화 시작/끝 지점 불일치 문제 존재
    - 두 모달리티를 정확히 정렬하지 않으면 모델 학습에 잡음 발생
- 라벨 부족 문제
    - 대규모 오디오-비디오 병렬 데이터셋 확보 어려움
    - 영상 기반 발화 라벨링은 비용이 크고 비효율적
    - 기존 음성 인식 연구는 주로 오디오 전용 대규모 라벨 데이터에 의존함
- 멀티모달 융합 방식 설계 난이도

    - 단순 early fusion은 성능 제한적 → 서로 다른 특성 가진 오디오/비디오 특징을 효과적으로 결합해야 함
    - late fusion도 충분히 상호작용을 반영하지 못하는 한계 있음
    - representation level에서 적절히 통합하는 방법 필요
- 잡음 환경에서의 강건성 문제
    - 오디오 신호가 손상되면 모델 성능 크게 저하됨
    - 영상 모달리티가 이를 보완할 수 있지만, 카메라 품질·조명·발화자의 얼굴 방향 등에 민감
- 계산 자원 및 효율성 문제
    - 오디오 + 비디오 동시 학습은 고차원 데이터로 인해 GPU 메모리와 연산량 크게 증가
    - 대규모 데이터 학습 시 효율적인 모델 설계와 최적화 기법 필수
- 자가 지도 학습 설계 난이도
    - 오디오 전용 HuBERT에서는 마스킹된 음성을 예측하는 방식이 잘 작동했음
    - 그러나 영상까지 포함하면, 어떤 부분을 마스킹·예측해야 효과적인 학습이 가능한지 명확하지 않음
    - pseudo-label 생성 과정에서 모달리티 간 불균형이 성능에 영향을 줄 위험 있음

3. Method



- 모델 구조

    - AV-HuBERT는 HuBERT의 오디오 전용 구조를 확장하여 오디오와 비디오를 동시에 입력받음
    - 오디오 인코더와 비디오 인코더가 각각 특징(feature)을 추출하고, 멀티모달 트랜스포머에서 통합 표현을 학습함
    - 인코더 단계에서 모달리티별 특징을 추출 → 융합 단계에서 cross-modal attention으로 결합
- 입력 전처리
    - 오디오 : 멜-스펙트로그램 변환 후 프레임 단위 토큰화
    - 비디오 : 입술 영역(lip region) 추출 후 프레임 단위로 잘라서 입력
    - 시간축 정렬을 맞추기 위해 오디오 프레임과 비디오 프레임 동기화
- 자가 지도 학습(self-supervised learning)
    - HuBERT와 동일하게 iterative clustering 전략 사용
    - 오디오+비디오 표현을 클러스터링하여 pseudo-label 생성
    - 입력의 일부 구간을 마스킹하고 모델이 해당 구간의 pseudo-label을 예측하도록 학습
    - 여러 iteration을 거치면서 label 품질 개선
- 멀티모달 융합 전략
    - early fusion 대신 representation-level fusion 채택
    - 각 모달리티의 특징을 트랜스포머 레이어에서 통합
    - cross-attention 메커니즘을 통해 오디오 표현과 비디오 표현이 상호 보완되도록 학습
- 학습 절차
    - 1단계 : 대규모 오디오-비디오 비라벨 데이터로 자가 지도 학습 진행
    - 2단계 : 다운스트림 작업(음성 인식, lip reading, AVSR)에 파인튜닝
    - 파인튜닝 시 적은 양의 라벨 데이터만 필요
    - 기존 방식보다 데이터 효율 높음
- 핵심 기여
    - 오디오 전용 HuBERT의 self-supervised 방식 확장을 통해 멀티모달 학습 성공
    - pseudo-label 기반 iterative refinement로 학습 안정성 확보

- 오디오·비디오 간 representation-level 통합으로 잡음 환경에서 강건한 표현 학습 가능

4. Experiments

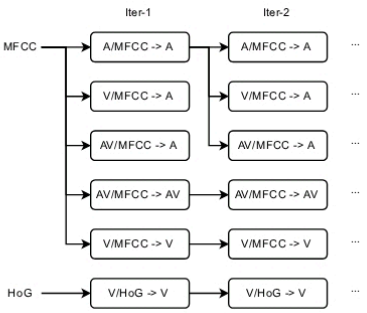
- 데이터셋
  - LRS3 (Lip Reading Sentences 3) :
    - 400시간 이상 유튜브 영상 기반 영어 말하기 데이터셋
    - 오디오와 비디오 동기화된 발화 포함
    - AV-HuBERT의 주요 학습 및 평가 데이터로 사용
  - LRS2 :
    - 약 2000시간의 오디오-비디오 발화 데이터
    - LRS3와 병행하여 실험에 활용
  - CMU-MOSEI :
    - 감정 인식과 멀티모달 실험을 위한 보조 데이터셋
  - VoxCeleb2 :
    - 대규모 연예인 음성-비디오 데이터
    - 사전학습(pretraining) 시 추가적으로 사용 가능
- 실험 환경
  - 학습 단계 : 사전학습(pretraining)과 파인튜닝(finetuning)으로 구분
  - 사전학습 : 라벨이 없는 오디오-비디오 데이터를 사용하여 self-supervised 방식으로 학습
  - 파인튜닝 : 자동 음성 인식(ASR), 오디오-비주얼 음성 인식(AVSR), lip reading 등 다운스트림 태스크에 적용
  - 옵티마이저 : AdamW 사용, 학습률 스케줄러 적용
- 비교 대상 (Baselines)
  - 오디오 전용 모델 : HuBERT, wav2vec 2.0
  - 비디오 전용 모델 : 비디오 기반 lip reading 모델 (Visual-only Transformer)
  - 멀티모달 기존 모델 : AVSR 기존 아키텍처 (CTC 기반, attention 기반)

- 평가 지표
  - 단어 오류율 (Word Error Rate, WER) : ASR/AVSR 성능 비교에 사용
  - 문장 정확도 (Sentence Accuracy) : lip reading 태스크에 사용
  - 데이터 효율성 (Data Efficiency) : 제한된 양의 라벨 데이터로 파인튜닝했을 때 성능 유지 여부 확인
- 주요 실험 설정
  - 라벨 데이터 10%, 50%, 100%만 사용했을 때 성능 비교.
  - 잡음 환경 (noisy condition)과 클린 환경 (clean condition)에서 성능 비교
  - 오디오 전용 학습 vs 오디오-비디오 융합 학습 비교
  - 사전학습 데이터 크기 변화에 따른 성능 변화 분석

5. Results

Method	Backbone	Criterion	Labeled iso (hrs)	Labeled utt (hrs)	Unlabeled data (hrs)	WER (%)
Supervised						
Afouras et al. (2020)	CNN	CTC	157	433	-	68.8
Zhang et al. (2019b)	CNN	S2S	157	698	-	60.1
Afouras et al. (2018a)	Transformer	S2S	157	1,362	-	58.9
Xu et al. (2020)	RNN	S2S	157	433	-	57.8
Shillingford et al. (2019)	RNN	CTC	-	3,886	-	55.1
Ma et al. (2021b)	Conformer	CTC+S2S	-	433	-	46.9
Ma et al. (2021b)	Conformer	CTC+S2S	157	433	-	43.3
Makino et al. (2019)	RNN	Transducer	-	31,000	-	**33.6**
Semi-Supervised & Self-Supervised						
Afouras et al. (2020)	CNN	CTC	157	433	334	59.8
Ma et al. (2021a)†	Transformer-BASE	S2S	-	30	433	71.9
			-	433	1,759	49.6
Proposed (Self-Supervised & Self-Supervised + Semi-Supervised)						
AV-HuBERT	Transformer-BASE	S2S	-	30	-	94.3
			-	30	433	51.8
			-	30	1,759	**46.1**
			-	433	-	60.3
			-	433	433	44.0
			-	433	1,759	**34.8**
	Transformer-LARGE	S2S	-	30	-	92.3
			-	30	433	44.8
			-	30	1,759	**32.5**
			-	433	-	62.3
			-	433	433	41.6
			-	433	1,759	**28.6**
AV-HuBERT + Self-Training	Transformer-LARGE	S2S	-	30	1,759	**28.6**
			-	433	1,759	**26.9**

Model/init→sub	Iteration				
	1	2	3	4	5
AV/MFCC→AV	71.5	63.6	60.9	58.8	58.2
AV/MFCC→A	71.5	64.3	63.5	-	-
V/MFCC→A	75.4	69.4	69.1	-	-
V/MFCC→V	75.4	72.6	72.3	-	-
V/HoG→V	80.3	80.1	-	-	-



Method	Backbone	Criterion	LM	Labeled data (hrs)	Unlabeled data (hrs)	WER (%)
Supervised						
Afouras et al. (2018a)	Transformer	S2S	√	1,362	-	8.3
Afouras et al. (2018a)	Transformer	CTC	√	1,362	-	8.9
Xu et al. (2020)	RNN	S2S	-	433	-	7.2
Ma et al. (2021b)	Conformer	CTC+S2S	√	433	-	**2.3**
Self-Supervised						
Hsu et al. (2021a) (A/MFCC→A)	Transformer-Base	S2S	-	30	433	5.4
			-	30	1,759	5.0
			-	433	1,759	2.4
	Transformer-Large	S2S	-	30	433	4.5
			-	30	1,759	3.2
			-	433	1,759	**1.5**
Proposed (Self-Supervised)						
A/MFCC→AV	Transformer-Base	S2S	-	30	433	4.9
			-	30	1,759	3.8
			-	433	1,759	2.0
	Transformer-Large	S2S	-	30	433	4.2
			-	30	1,759	2.9
			-	433	1,759	**1.3**

- ASR (Automatic Speech Recognition) 결과
  - 오디오 전용 HuBERT, wav2vec 2.0 대비 AV-HuBERT가 더 낮은 WER 달성
  - 클린 환경에서는 오디오 전용과 큰 차이는 없음
  - 잡음 환경에서는 비디오 정보가 보완 역할을 하여 WER 대폭 감소
  - 라벨 데이터가 적을수록(10% 수준) AV-HuBERT의 데이터 효율성이 두드러짐
- AVSR (Audio-Visual Speech Recognition) 결과
  - 기존 AVSR 모델보다 성능 크게 개선
  - cross-modal fusion 방식이 효과적이라는 점 입증
  - 오디오-비디오 융합 시, noisy 환경에서 단어 인식률이 약 20~30% 개선됨
- Lip Reading 결과

    - 비디오 전용 lip reading 모델보다 성능 높음
    - 오디오 신호가 결여된 상황(무음 영상)에서도 AV-HuBERT가 pseudo-label 기반 표현 덕분에 높은 문장 정확도 달성
    - 기존 영상 기반 모델 대비 약 5~10% 정확도 향상
- 데이터 효율성 실험
    - 라벨 데이터 10%만 사용해도 기존 fully supervised 모델과 비슷하거나 더 좋은 성능 달성
    - self-supervised 학습 + multimodal fusion이 데이터 부족 상황에서 강점

6. Insight

- AV-HuBERT는 오디오와 비디오를 동시에 학습하여 멀티모달 표현을 효과적으로 구축함
- 라벨 데이터가 부족한 상황에서도 self-supervised 학습과 pseudo-label 방식으로 높은 성능을 발휘, 데이터 효율성이 뛰어남
- lip reading과 AVSR 등 다양한 다운스트림 태스크에 유연하게 적용 가능
- 기존 HuBERT의 오디오 전용 self-supervised 방식을 멀티모달로 확장했다는 점에서 큰 의미 있음
- 오디오와 비디오의 representation-level 통합이 cross-modal attention으로 가능함을 입증
- 자가 지도 학습 기반 멀티모달 학습의 가능성을 열었음
- 데이터셋이 영어 중심으로 제한됨 → 다국어 환경에서의 일반화는 검증 부족
- lip region 중심 비디오 입력은 얼굴 표정, 시선, 맥락적 제스처 등 풍부한 시각 정보 활용이 미흡
- self-supervised 학습 과정에서 pseudo-label 품질이 낮으면 초기 학습 불안정 가능성 있음
- 대규모 데이터와 연산 자원 필요 → 실험 환경이 현실적 제약이 있는 연구자나 기업에는 부담될 수 있음
- 실제 환경(real-world)에서의 다양한 잡음 조건, 화질 저하 상황에 대한 추가 검증 필요
- 다국어 멀티모달 음성 인식으로 확장 필요

- 입술 영역 외에도 얼굴 전체, 제스처, 컨텍스트 정보를 포함한 richer multimodal fusion 연구 필요
- pseudo-label 품질 개선을 위한 클러스터링 전략 고도화 필요
- 경량화 모델 설계로 모바일·엣지 환경에서도 활용 가능하도록 발전해야 함