ASAC 7기 DL 1조

Speech Recognition

멀티모달 학습을 활용한 음성 인식

1조 : 김선규 박건 유민지 이승연 장길만

Contents

멀티모달 학습을 활용한 음성 인식

AVSR 연구 동향과 소음 환경에서의 성능 한계

최근 연구 동향 및 한계

01

- **AVSR(Audio-Visual Speech Recognition) 개발 활발**
- **기존 STT & ASR보다 멀티모달(오디오+비주얼)의 강력한 성능**

02

- **하지만 실제 영상에서 입이 가려지거나 비디오 품질 저하 시 성능 저하 발생**
- **소음 환경에서 여전히 약점 존재**

SST & ASR vs AVSR 차이점

STT → ASR → AVSR

STT & ASR
입력 데이터 : 오디오

- **STT (Speech to Text)**

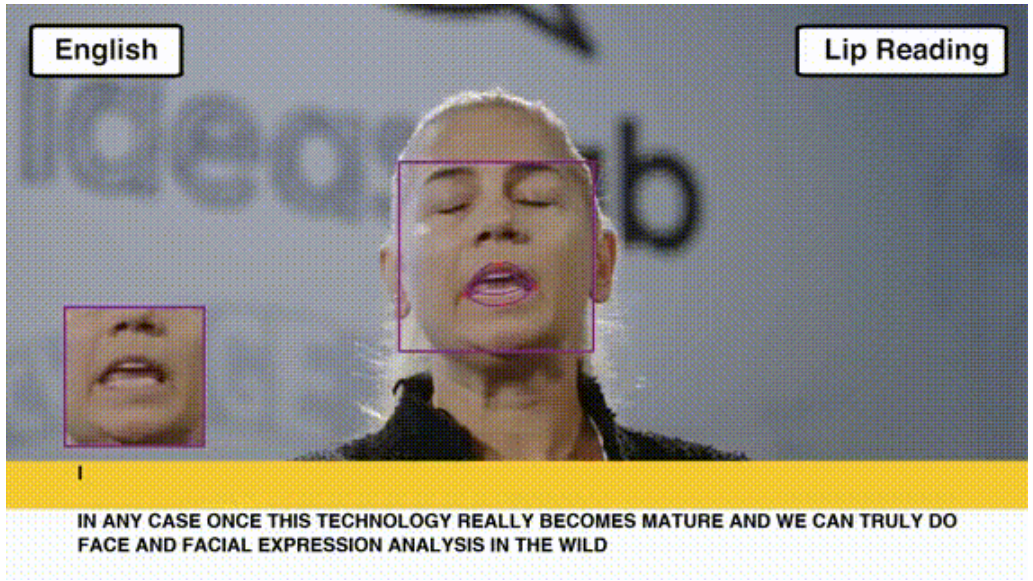
  음성 데이터를 텍스트로 변환하는 기술

- **ASR (Automatic Speech Recognition)**

  STT를 포함하며, 자동으로 문맥을 이해하여 문장 형성하는
  음성 인식 시스템

SST & ASR vs AVSR 차이점

STT → ASR → AVSR



AVSR
입력 데이터 : 오디오 + 비디오 (입 모양)

- **AVSR (Audio-Visual Speech Recognition)**
  - 오디오와 **입 모양 (비디오) 정보까지** 활용하여 음성 인식
  - 오디오와 비디오 모두 활용하여 **실시간 자막의 정확도 극대화**
  - **멀티모달 학습** (오디오 + 비디오)로 인식률 향상

  - 주요 기술 : **AV-HuBert, USR 등**

실시간 자막의 필요성 – STT & ASR 한계

기존 STT/ASR은
배경 소음이 많거나, 발화자의 발음이 명확하지 않을 때 성능 저하 발생

모델 유형	입력 데이터	주요 한계점
STT (Speech-to-Text)	오디오	주변 소음에 취약
ASR (Automatic Speech Recognition)	오디오 + 문맥 기반 분석	오디오 품질 저하 시 성능 저하
VSR (Visual Speech Recognition)	입술 움직임(비디오)	오디오 없이 인식 가능하지만, 단독 사용 시 한계
AVSR (Audio-Visual Speech Recognition)	오디오 + 비디오(입 모양)	오디오, 비디오 정보 모두 활용 소음 환경에서도 강력한 성능 유지

1. 서론 및 배경

한국어 립리딩이 아닌 영어 모델을 사용한 이유

왜 한국어가 아닌, 영어를 선택했는가?

- 한국어 립리딩은 데이터 부족, 언어적 차이, 기술적 한계로 인해 연구가 활발하지 **않음**.

- 기존 영어 모델 적용이 어려워 **한국어 특화된 접근 방식**이 필요함

데이터 부족

- 대규모 한국어 립리딩 데이터셋 부재

- 기존 연구: 영국 BBC 방송 영상 데이터 사용

언어적 차이

- 영어는 26개의 알파벳, 한국어는 **11,172개의 음절 조합**

- 초성, 중성, 종성 단위의 새로운 립리딩 접근 방식 필요

+

기술적 한계

- 기존 알고리즘은 영어 중심 설계로 **한국어 적용 어려움**

- 학습되지 않은 단어 등장 시 **인식이 어려움**

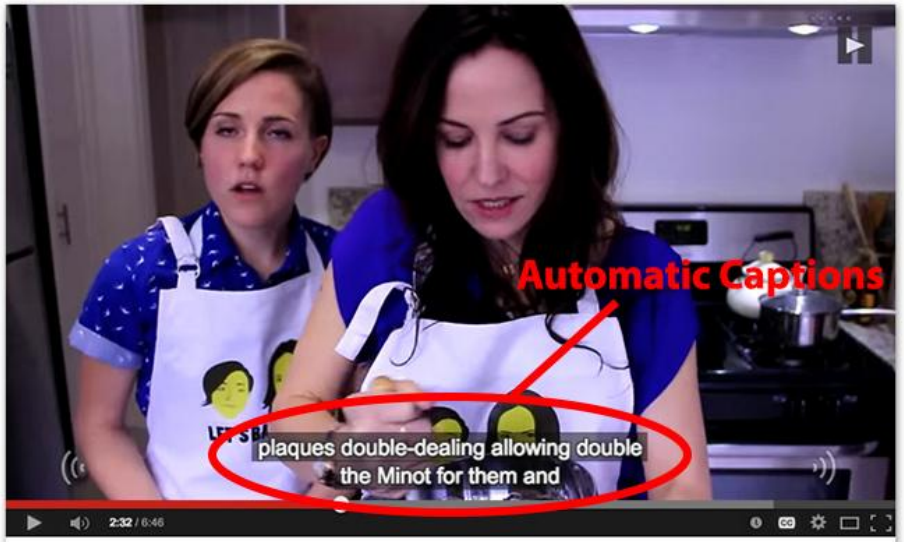
제한된 연구

- 한국어 립리딩을 지원하는 오픈소스 및 사전 학습 모델이 **거의 없음**

* 출처 : 조선영, 윤수성, 「한국어 립리딩: 데이터 구축 및 문장수준 립리딩」, Journal of the KIMST, Vol. 27, No. 2, 국방과학연구소, 2024, 167-176.

기존 음성 인식 한계

STT (음성 –기반 자동 자막 생성) 유튜브 실패 사례



- 기존 ASR (음성 기반 인식) 한계

- 배경 소음이 있으면 음성 인식률 저하
- 화자의 입 모양을 고려하지 않아 정확도 낮음

"기존 STT의 한계를 극복하기 위해
오디오와 비디오 정보를 활용하는
AVSR 모델 적용이 필요"

* AVSR : 음성과 입 모양을 함께 학습하여
소음환경에서도 높은 정확도 유지

기존 모델 한계점

기존 모델들은 멀티모달(음성+비디오) 기반이지만,
소음 환경에서 여전히 취약

모델	특징	학습 방식	한계점
Visual Speech Recognition (VSR)	입술 모양 및 움직임만 보고 텍스트로 변환	지도 학습 (입술 영상만 사용)	음성없이 입술 모양만 보고 해석
Whisper Flamingo	Whisper(음성인식) + Flamingo (영상정보) 결합	지도 학습 (음성과 입 모양 및 얼굴 움직임 비주얼 데이터 사용)	오디오가 왜곡되면 보완 불가
Auto-AVSR	오디오 + 비디오 동시 학습	반지도 학습 (Pseudo-Labeling)	입 모양 정보 부족 시, 성능 감소
USR (Unified Speech Recognition)	오디오 + 비디오 동시 학습	반지도 학습 (Pseudo-Labeling + 멀티모달 특징 추출)	최신 연구 모델

2. 기존 연구 및 최신 모델 비교

Auto-AVSR 모델 구조

AUTO-AVSR : "ASR 모델을 활용한 자동 라벨링"

공개된 사전학습 ASR 모델 (Whisper, wav2vec2, HuBERT 등)을 이용하여 라벨이 없는 영상 데이터(AVSpeech, VoxCeleb2)에 자동으로 라벨을 생성

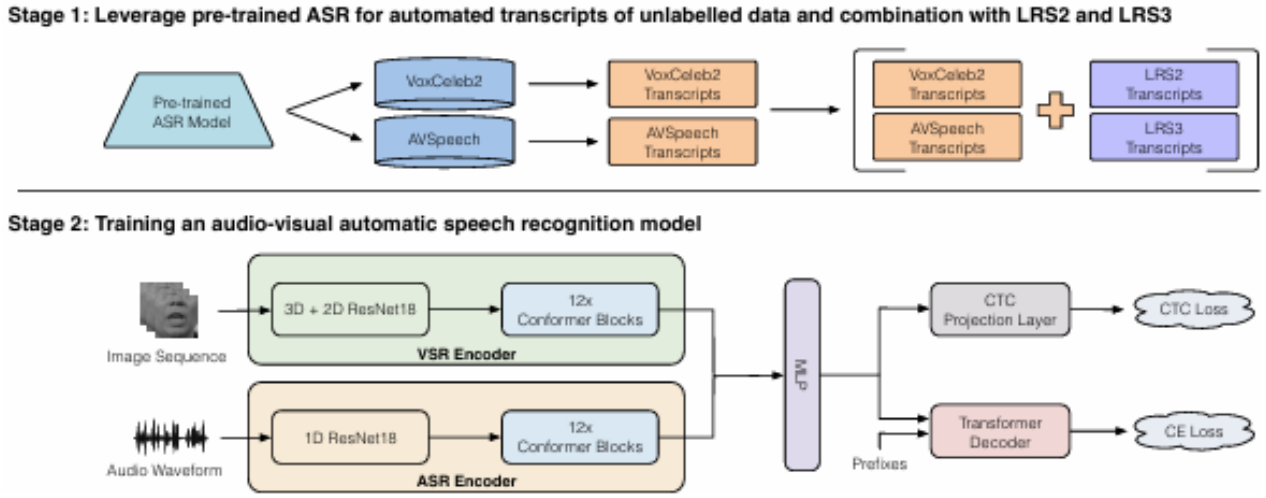


Fig. 1. AV-ASR architecture overview. In the first stage, a pre-trained ASR model is leveraged to produce automatically-generated transcriptions for unlabelled audio-visual datasets. And then these unlabelled datasets are combined with the labelled training sets, including LRS2 and LRS3, for training. The frame rate of audio and visual features from the ASR and VSR encoders is 25 frames per second (fps).

Stage 1: 자동 Mapping 생성해 학습 데이터 확장

Stage 2: 오디오와 비디오를 함께 사용하는 AV-ASR 모델을 학습

* 출처 : P. Ma, A. Haliasos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "AUTO-AVSR: Audio-Visual Speech Recognition with Automatic Labels," Imperial College London & Meta AI, UK.

2. 기존 연구 및 최신 모델 비교

정확도 및 WER 계산식

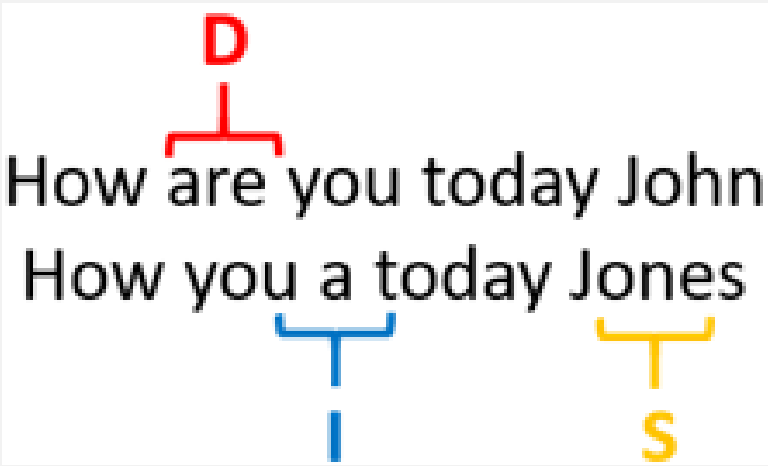
WER(Word Error Rate)

- 정답(Reference) : How are you today John
- 예측 가설(Hypothesis) : How you a today Jones

$$WER = \frac{S+D+I}{N}$$

단어 수(N) : 5
대체 오류(S) : 1 (John -> Jones)
삽입 오류(I) : 1 (you today -> you a today)
삭제 오류(D) : 1 (How are you -> How you)

WER : 단어 오류율



WER = 1 + 1 + 1 / 5 = 60%
정확도 = 1 – WER = 40%

2. 기존 연구 및 최신 모델 비교

SNR (신호 대 잡음 비율)

SNR(Signal-to-Noise Ratio)

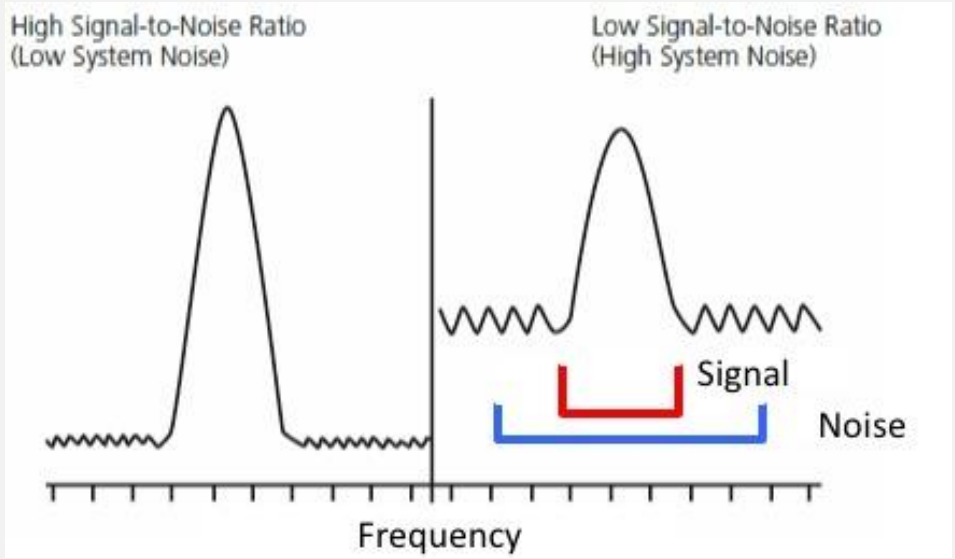
SNR(dB): "원하는 신호와 노이즈의 강도를 비교"하는 척도

$$\mathrm{SNR\ (dB)} = 10 \log_{10} \left(\frac{P_s}{P_n} \right)$$

일반적으로 데시벨(dB)로 표현 되며, 위의 식과 같음

- **원하는 신호의 평균 전력 : Ps**
- **노이즈 신호의 평균 전력 : Pn**

SNR : 신호 대 잡음 비율



* 출처 : https://audio-engineer-lounge.blogspot.com/2015/02/thd-snr.html

2. 기존 연구 및 최신 모델 비교

인위적인 소음이 있을 때와 없을 때 AVSR vs ASR vs VSR 성능 비교

*** 소음 없는 BBC 영상 성능 비교**

모델	정확도	WER
AVSR	95.74%	4.26%
ASR	95.74%	4.26%
VSR	89.36%	10.64%

*** 소음 있는 BBC 영상 성능 비교 (Babble Noise -5dB 환경)**

모델	정확도	WER
AVSR	93.62%	6.38%
ASR	91.49%	8.51%
VSR	91.49%	8.51%

2. 기존 연구 및 최신 모델 비교

소음이 없는 영상 : AVSR vs ASR vs VSR 성능 비교

*** 소음 없는 BBC 영상 성능 비교**

모델	정확도	WER
AVSR	95.74%	4.26%
ASR	95.74%	4.26%
VSR	89.36%	10.64%

*** 문장 비교**

정답	release of the names of those three israeli hostages who are due to be freed tomorrow. At one point in our coverage we mistakenly called the hostages prisoners and we would like to apologise for that error
AVSR	release of the names of those three israeli hostages who are due to be freed tomorrow at one point in our coverage we mistakenly call the hostages prisoners and we would like to apologize for that
ASR	release of the names of those three israeli hostages who are due to be free tomorrow at one point in our coverage we mistakenly call the hostages prisoners and we would like to apologize for that
VSR	release on the names of those three israeli hostages who are known to be free tomorrow at one point in our coverage we mistakenly called hostages **and** prisoners and we would like to apologize for that error

2. 기존 연구 및 최신 모델 비교

소음이 있는 영상 : AVSR vs ASR vs VSR 성능 비교 – Babble Noise 5dB

*** 소음 있는 BBC 영상 성능 비교**

모델	정확도	WER
AVSR	95.74%	4.26%
ASR	95.74%	4.26%
VSR	89.36%	10.64%

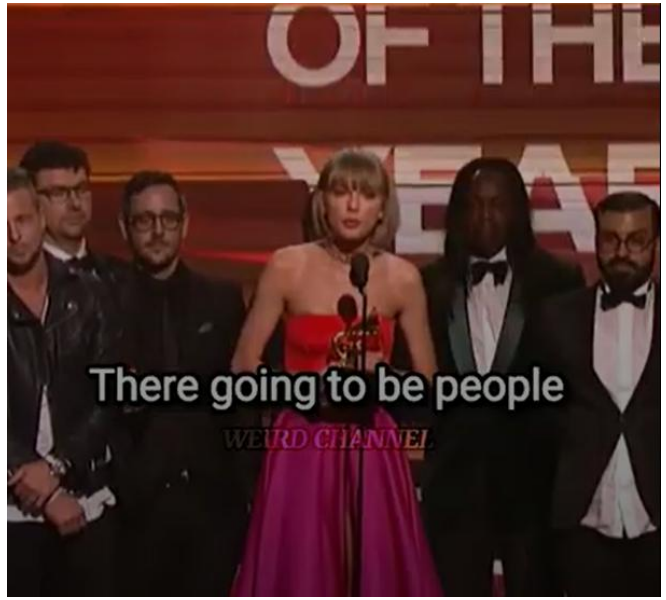
*** 문장 비교**

정답	who are **due** to be freed tomorrow. At one point in our coverage we mistakenly called the hostages prisoners and we would like to apologize for that error
AVSR	who are due to be freed tomorrow at one point in our coverage we mistakenly call the hostages **to** prisoners and we would like to apologize for that
ASR	who are due to be free tomorrow at one point in our coverage we mistakenly call the hostages **of** prisoners and they would like to apologize for that error
VSR	who are known to be free tomorrow at one point in our coverage we mistakenly called hostages imprisoned and we would like to apologize for that error

2. 기존 연구 및 최신 모델 비교

AVSR 성능

* 복수 인물 발화 환경에서 성능 비교



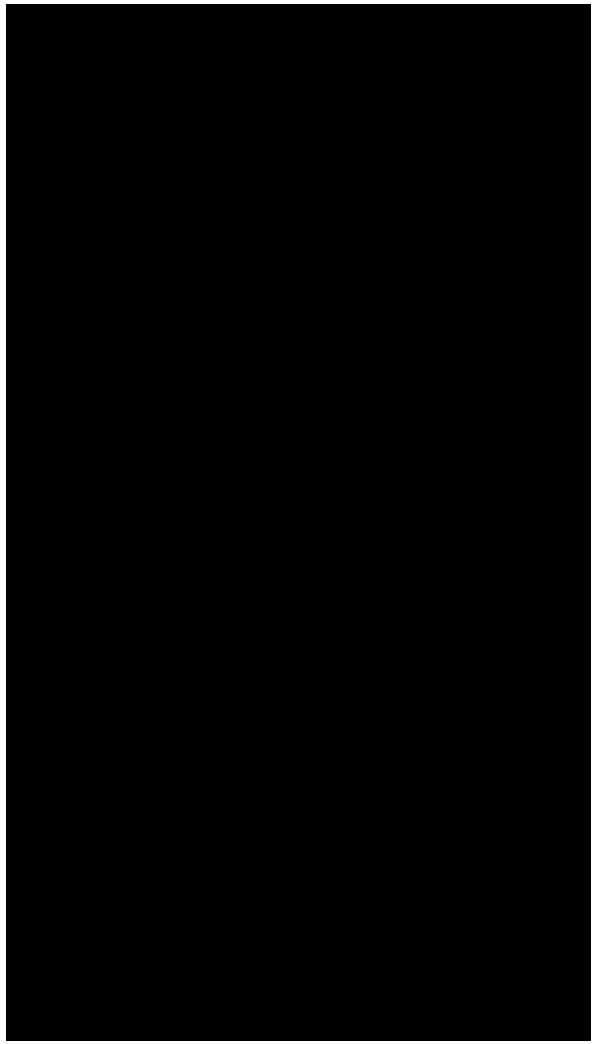
* AVSR

모델	정확도 (%)	WER (%)
AVSR	28	72
VSR	-48	148
ASR	48	52

-> AVSR : 사람이 많거나 입 모양 감지 불가할 시 성능이 현저히 떨어짐

2. 기존 연구 및 최신 모델 비교

AVSR vs ASR vs VSR 성능 비교

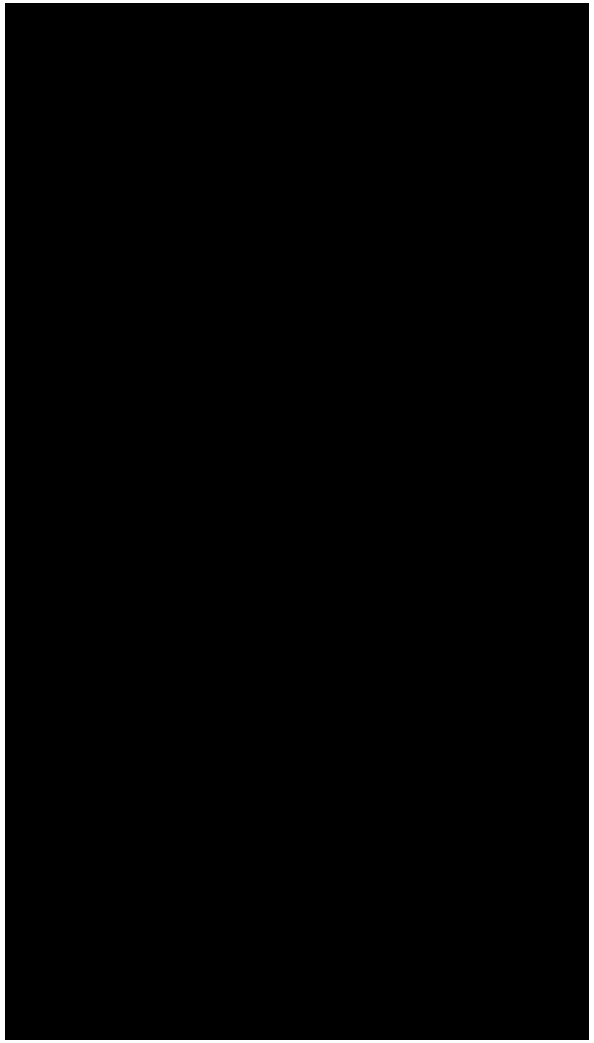


*** 문장 비교**

정답	I wanna say to all The young woman out there There going to be people along the way who will try to take undercut your success or take credit for your accomplishments or your fame but if you just focus on the work and you don't let those people sidetrack you someday when you get where you're going you'll look around and you will know that it was you and the people who love you who put you there and that will be the greatest feeling in the world thank you for this moment

2. 기존 연구 및 최신 모델 비교

AVSR vs ASR vs VSR 성능 비교



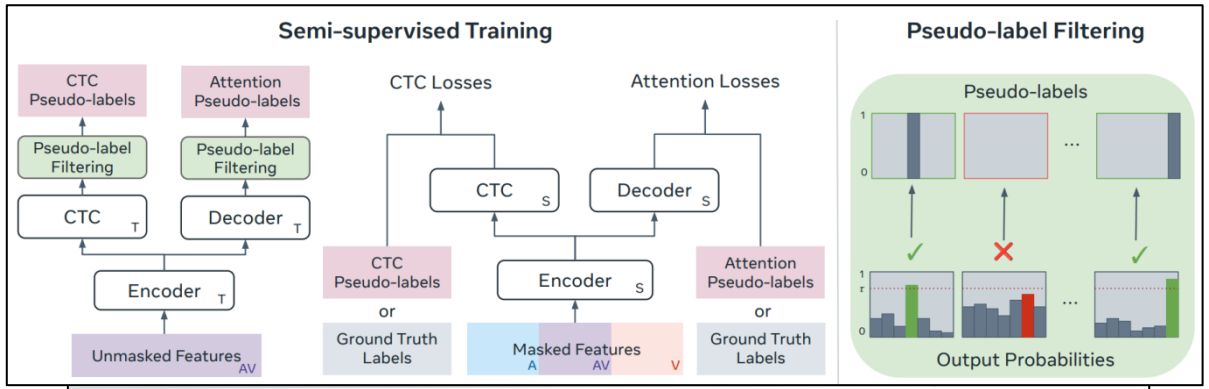
*** 문장 비교**

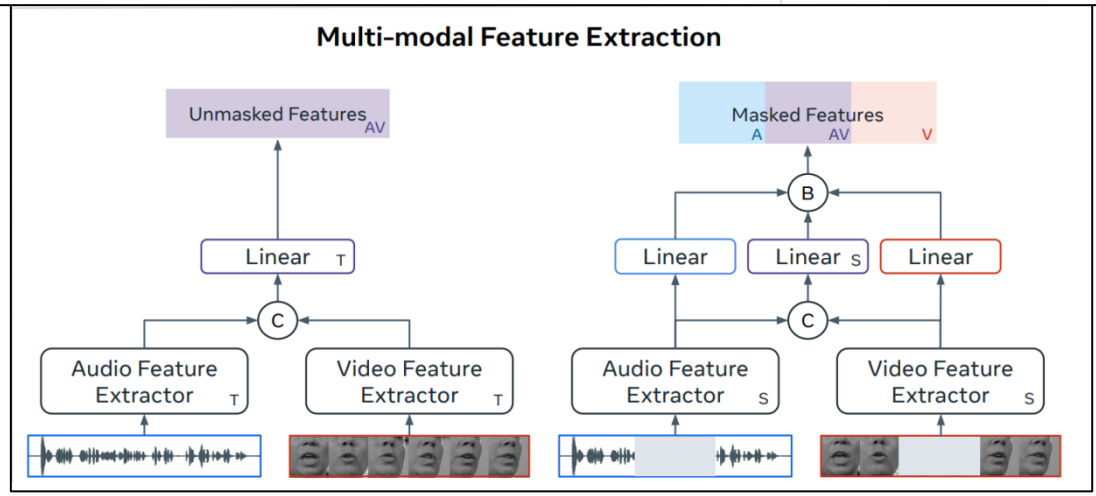
AVSR	I want to say to all the young women out there There are going to be people along the way who will try to take undercut your success or take credit for your accomplishments or your fame but if you just focus on the work and you don't let those people sidetrack you someday when you get where you're going you'll look around and that will be the greatest feeling in the world thank you for this moment (반복)**thank you for this moment and that's what we're going to do we'll look around and you for this moment thank you thank you for this moment**
ASR	i want to say to all the young women out there there are going to be people along the way who will try to take undercut your success or take credit for your accomplishments or your fame but if you just focus on the ~~fact that you're not going to be able to do that i think~~ you're going ~~to be able to do~~ that ~~i think you're going to be able to do~~ that ~~i think you're going to~~ be ~~a part of~~ the world
VSR	AND WHEN YOU SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEEYOU'LL SEE YOU'LL SEE YOU'LL SEE ....(반복)

USR (Unified Speech Recognition)

" USR 모델 채택 이유 :
다중 발화자 환경에서도 안정적인 성능"

- "반지도 학습"
- "Pseudo-Label Filtering" : 다중 화자에서도 성능 유지
- "Multi-Modeal Feature" : 오디오와 비디오 정보를 결합





3. 모델 학습 로직 및 성능 비교

USR 소음에 대한 성능 비교

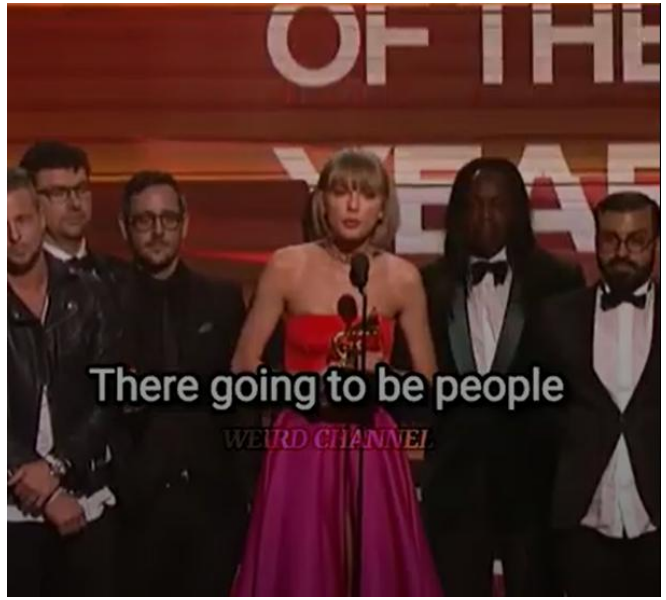
* 소음이 음수로 가까워지면 (증가할 수록)
-> 성능이 저하됨

소음환경	모델	정확도 (%)	WER (%)
원본	AV	69.23	30.77
	A	80.77	19.23
	V	53.85	46.15
소음 (-5dB)	AV	-26.92	126.92
	A	-11.54	111.54
	V	26.92	73.08
소음 (0dB)	AV	19.23	80.77
	A	19.23	80.77
	V	53.85	46.15
소음 (+5dB)	AV	69.23	30.77
	A	73.08	26.92
	V	30.77	69.23

3. 모델 학습 로직 및 성능 비교

USR 성능 한계

* 복수 인물 발화 환경에서 성능 비교



* USR

모델	정확도 (%)	WER (%)
AV	68	32
A	84	16
V	40	60

* AVSR

모델	정확도 (%)	WER (%)
AVSR	28	72
VSR	-48	148
ASR	48	52

-> USR : AVSR 보다 전체 점수(정확도 및 WER)가 높지만
소음환경과 사람 입모양 감지가 어려울 시, 여전히 낮은 점수를 보임

USR (Unified Speech Recognition)

- 입모양 크롭 기술(MediaPipe)와 결합 시

다중 발화자 환경에서 인식 정확도 높음

- 다중 화자에서도 성능 저하 없이 정확도 유지

3. 모델 학습 로직 및 성능 비교

USR vs AVSR 성능 비교

*** USR**

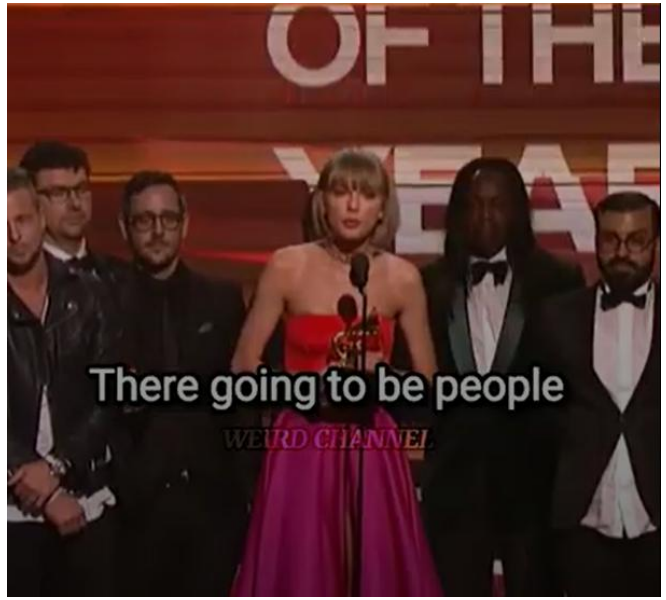
모델	정확도 (%)	WER (%)
AV	69.23	30.77
A	80.77	19.23
V	53.85	46.15

*** AVSR**

모델	정확도 (%)	WER (%)
AVSR	100	0
ASR	96.15	3.85
VSR	100	0

3. 모델 학습 로직 및 성능 비교

USR vs AVSR 성능 비교



*** USR**

모델	정확도 (%)	WER (%)
AV	68	32
A	84	16
V	40	60

*** AVSR**

모델	정확도 (%)	WER (%)
AVSR	28	72
VSR	-48	148
ASR	48	52

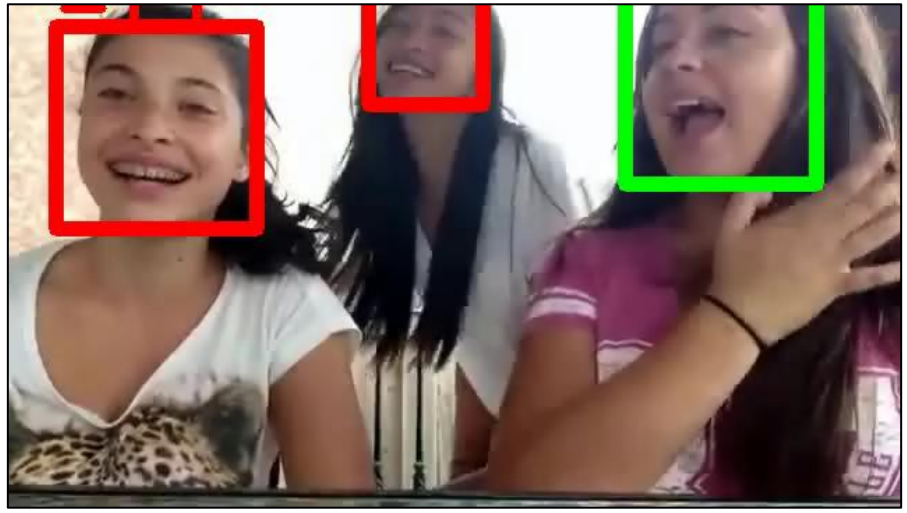
4. 실험 방법 및 진행 과정

음성-영상 기반 화자 인식

TalkNet : Audio-visual Active Speaker Detection

영상속에서 누가 실제로 말하고 있는지 탐지 셀프 어텐션 (self-attention)*으로 화자 활동을 시계열적으로 분석 후 추론

* 영상 전체 흐름을 고려해서, 현재 시점이 정말 "말하고 있는 순간"인지 정확하게 판단



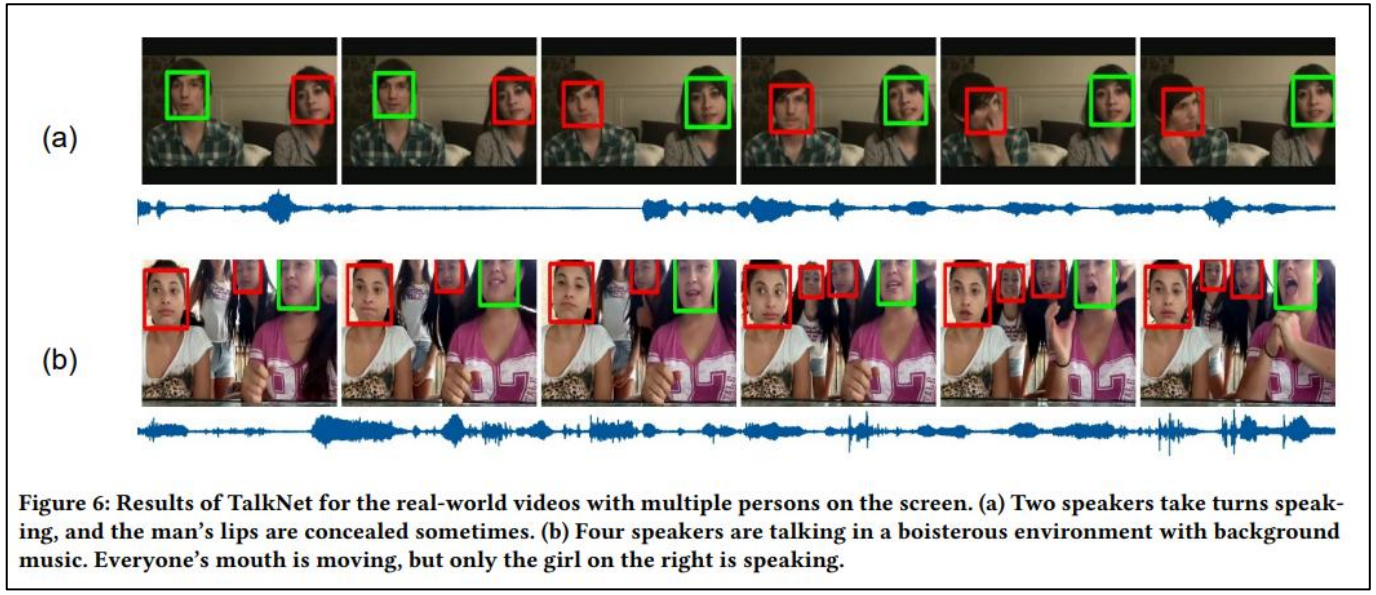


Figure 6: Results of TalkNet for the real-world videos with multiple persons on the screen. (a) Two speakers take turns speaking, and the man's lips are concealed sometimes. (b) Four speakers are talking in a boisterous environment with background music. Everyone's mouth is moving, but only the girl on the right is speaking.

출처 : Tao et al., Is Someone Speaking? Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection, ACM MM 2021.

4. 실험 방법 및 진행 과정

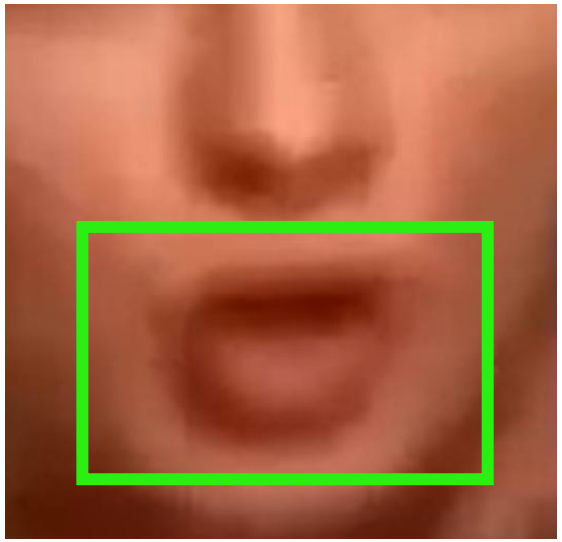
데이터 전처리 과정

* Librosa에서 사운드 추출

TalkNet

Media Pipe

* MediaPipe 기반 얼굴 랜드마크(특징점)
→ 입술 영역 좌표 기반 입 모양 추출
→ 전처리 데이터로 저장



librosa

파이썬 기반 오디오 신호 처리 라이브러리

MediaPipe

Google이 개발한 멀티모달 ML 파이프라인 프레임워크
실시간으로 영상/이미지 속 사람의 랜드마크 탐지

5. 연구 결론 및 향후 개선 방향

화자 감지 기능을 추가한 USR 성능 비교

*** 화자 감지 기능 추가한 USR**

모델	정확도 (%)	WER (%)
AV	84	16
A	100	0
V	16	84

*** USR**

모델	정확도 (%)	WER (%)
AV	68	32
A	96	4
V	40	60

5. 연구 결론 및 향후 개선 방향

AVSR과 USR: 발화자 탐지 여부에 따른 성능 비교

- **Auto-AVSR : 발화자 탐지 X**

YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEE YOU'LL SEEYOU'LL SEE YOU'LL SEE YOU'LL SEE ... (반복)

성능 향상

- **USR : 발화자 탐지 O**

or take credit for your accomplishments or your fame **ns** but if you just focus on the work and you don't let those people sidetrack you **that's why you**

감사합니다.

멀티모달 학습을 활용한 음성 인식