JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

THU-PM-221

AVFormer:
Injecting Vision into Frozen Speech Models for Zero-Shot AV-ASR

Paul Hongsuck Seo, Arsha Nagrani, Cordelia Schmid

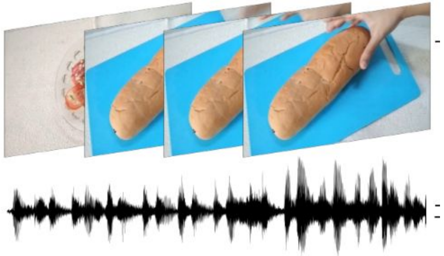
Google Research

Overview: **Task** is Audiovisual ASR (AV-ASR)

Incorporate **visual** context to ASR for **robust** speech recognition

- *Useful in cases of heavily accented speech, background noise, ambiguous pronunciation*
- *Goes beyond lip motion*
  - *Visual frames can provide clues of objects, actions, backgrounds*

Audiovisual input stream



*I am using a garlic **loaf** to make the sandwich*

Transcribe the speech, use the vision to help

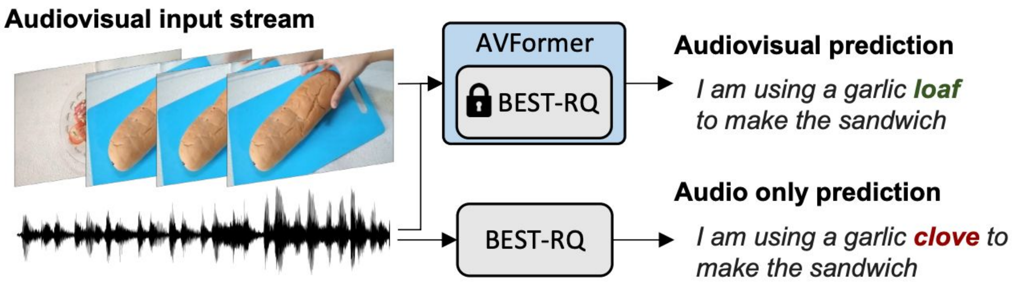
Google Research

Overview: **Method**

Existing AV-ASR methods are developed **from scratch on new benchmarks.**
Re-inventing the wheel?

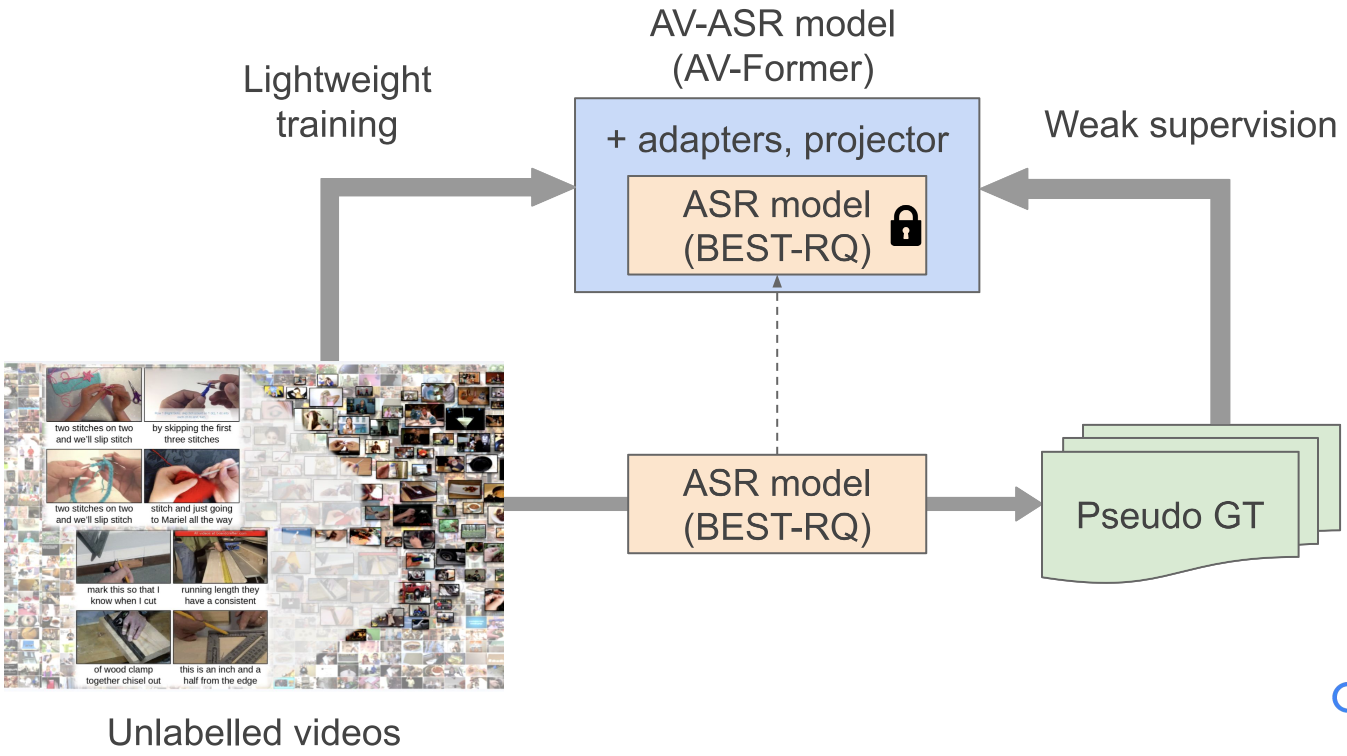
However,
there are many **strong, large-scale** ASR models (eg. Best-RQ), and visual models (CLIP)

*Can we simply **inject CLIP visual features** into highly-engineered **existing ASR models** **without a new annotated dataset**?*



Google Research

Overview: **Train** on pseudo labels from HowTo100M



Google Research

Overview: **Results**

State-of-the-art **zero-shot** results across **3 AV-ASR datasets**: spanning instructional videos and egocentric home videos

How2

Object "schooner" corrected to "scooter"

VisSpeech

Object "clove" corrected to "loaf"

Ego4D

"Drive" recognised from the road!

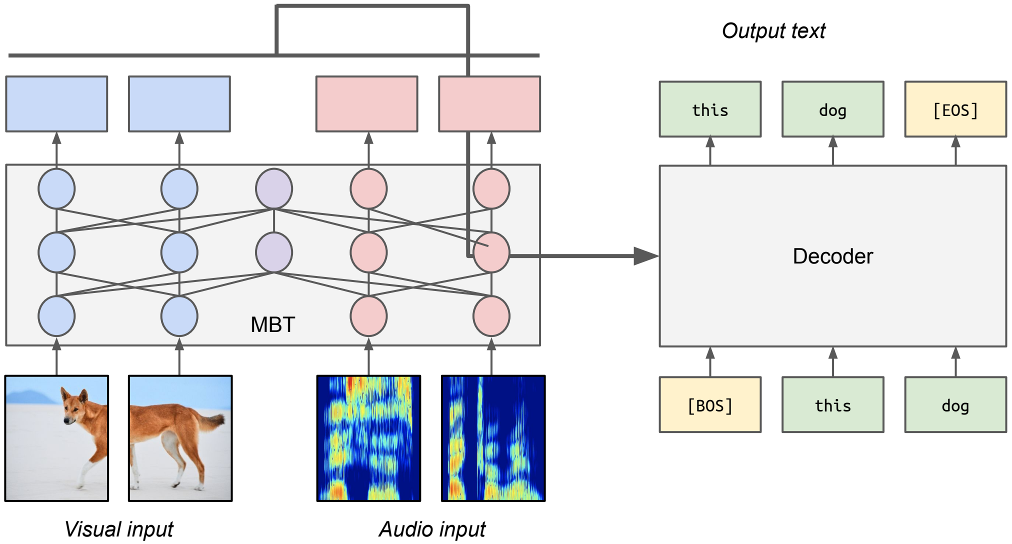
Google Research

Existing AV-ASR methods

Developed from **scratch** and trained end-to-end

Costly to train (early fusion of audio and video)

Do not **generalise** well to new domains



AVATAR: Unconstrained Audiovisual Speech Recognition, Gabeur et al. Interspeech 2022

Google Research

Existing AV-ASR methods

Developed from **scratch** and trained end-to-end

However,

there are many **strong, large-scale** ASR models (eg. [Best-RQ]) that are

- Huge (billions of params)
- Trained with self-supervision in the audio domain
- Generalize well to new domains
- Achieve amazing performance on audio-only ASR benchmarks

There are **strong vision** models (eg. [CLIP])

  - Also great generalization

Google Research

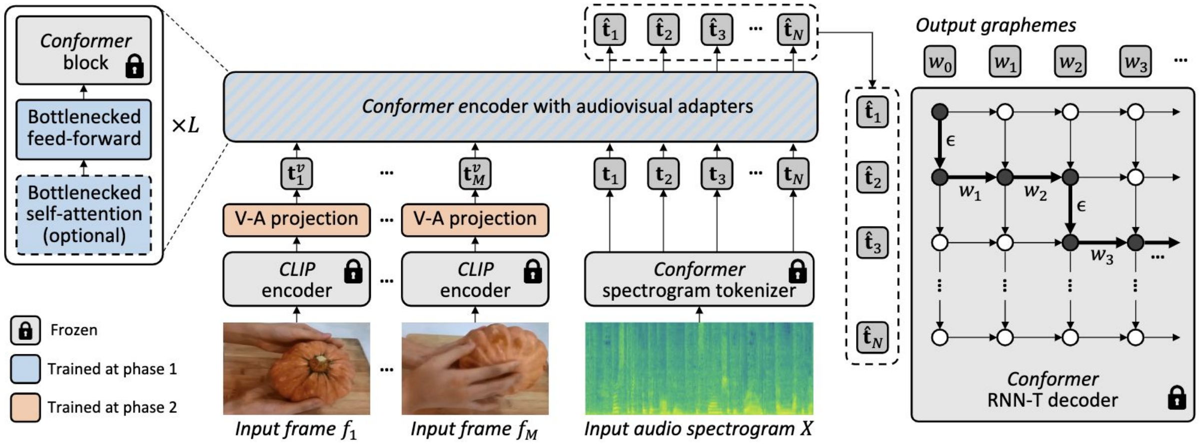
AVFormer: Lightweight Adapters

Method: Start with a frozen SOTA ASR model (Best-RQ, ICML'22-paper based on the conformer) and add frozen CLIP features to the input

Add **two** types of adapter layers

1) Bottleneck layers in the encoder block (allow domain adaptation)
2) Visual projection layers (that transform CLIP features)

Only Adapters are trained, rest is frozen



Google Research

AVFormer: Curriculum Strategy

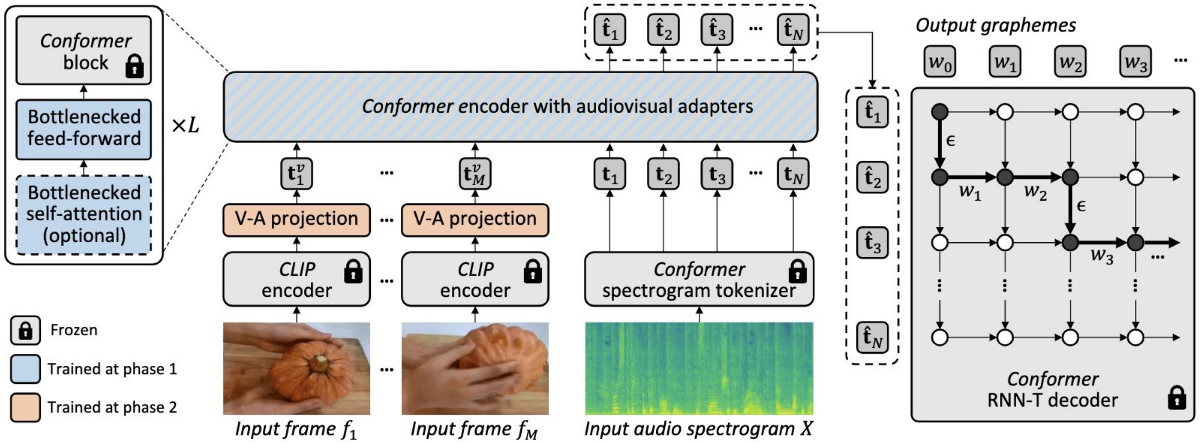
Method: Start with a frozen SOTA ASR model (Best-RQ, ICML'22-paper based on the conformer) and add frozen CLIP features to the input

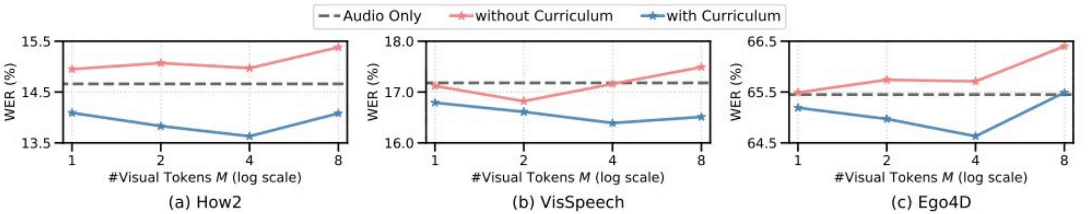
Add **two** types of adapter layers

1) Bottleneck layers in the encoder block (allow domain adaptation)
2) Visual projection layers (that transform CLIP features)

Only Adapters are trained, rest is frozen

Curriculum Strategy: Train (1) first then (2), crucial to allow model to use visual information





Lower WER is better, curriculum helps, visual tokens help

Train on pseudo labels from HowTo100M

Pre-training for Best-RQ (audio-only)

- LibriLight: large-scale unlabelled speech dataset.
- LibriSpeech: ASR benchmark with GT annotation but without visual inputs.

Lightweight training (non-transcribed videos with pseudo ground truth)

- HowTo100M: large-scale unannotated instructional videos.

Evaluate on zero-shot AV-ASR benchmarks

Pre-training for Best-RQ (audio-only)

- **LibriLight:** large-scale unlabelled speech dataset.
- **LibriSpeech:** ASR benchmark with GT annotation but without visual inputs.

Lightweight training (non-transcribed videos with pseudo ground truth)

- **HowTo100M:** large-scale unannotated instructional videos.

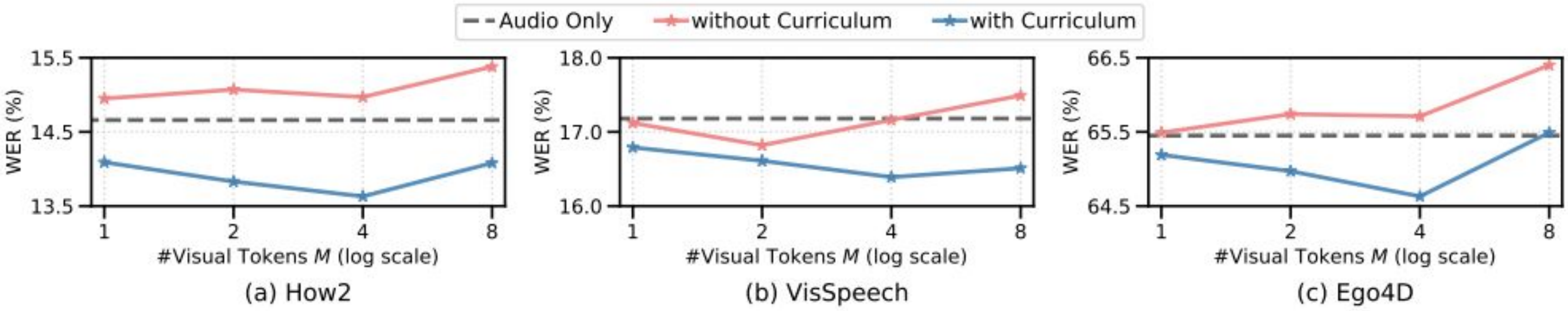
Evaluation (transcribed videos from different domains) zero-shot

- **How2:** Instructional video clips with pseudo-GT from user generated captions.
- **VisSpeech:** Instructional video clips with high speech-vision correlation with GT.
- **Ego4D:** Egocentric video clips with GT.

Results in WER, Lower is better

Ablations

The curriculum helps on all 3 zero-shot AV-ASR datasets

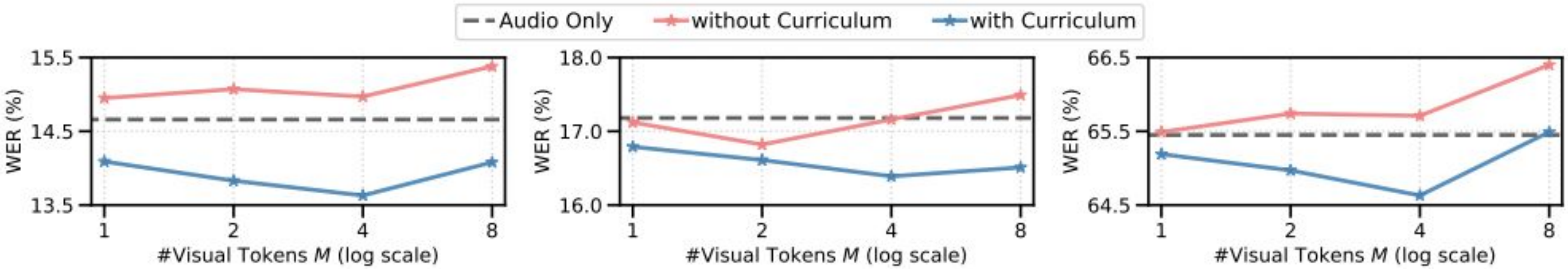


Google Research

Results in WER, Lower is better

Ablations

The curriculum helps on all 3 zero-shot AV-ASR datasets



(a) How2 (b) VisSpeech (c) Ego4D

The gains from the visual tokens and the adapters are complementary

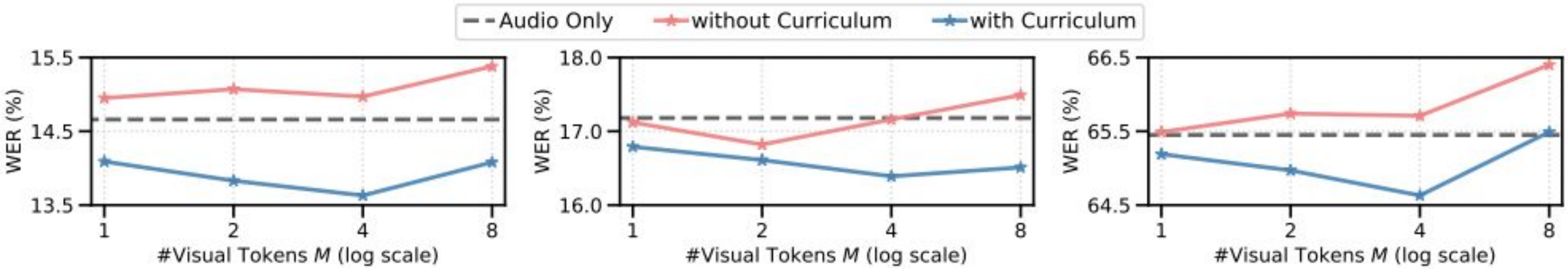
Visual tokens	Adapters	How2	VisSpeech	Ego4D
		21.90	31.61	77.98
✓		19.74	31.13	76.50
	✓	14.66	17.18	65.45
✓	✓	13.63	16.39	64.63

Google Research

Ablations

Results in WER, Lower is better

The curriculum helps on all 3 zero-shot AV-ASR datasets



(a) How2 (b) VisSpeech (c) Ego4D

The gains from the visual tokens and the adapters are complementary

Visual tokens	Adapters	How2	VisSpeech	Ego4D
		21.90	31.61	77.98
✓		19.74	31.13	76.50
	✓	14.66	17.18	65.45
✓	✓	13.63	16.39	64.63

*Only **5%** of the HowTo100M dataset is needed to train our **lightweight** adapters*

Training dataset size	How2	VisSpeech	Ego4D
5%	13.69	16.60	64.75
100%	13.63	16.39	64.63

Google Research

Comparison to SOTA

- SOTA on zero-shot AV-ASR benchmarks (How2, VisSpeech, Ego4D) **across domains**
- While **maintaining** performance on traditional ASR benchmarks (Librispeech)

Note: Zero-shot ASR is a more useful setting for *application/production*

Method	Modality	LibriSpeech PT	HowTo100M PT Pretrained params	HowTo100M PT Data %	LibriSpeech	How2	VisSpeech	Ego4D
AVATAR [11]	A	✓	–	–	8.85	39.43	65.33	110.86
AVATAR [11]	A+V	–	All	100	24.65	17.23	35.66	92.03
AVATAR [11]	A+V	✓	All	100	24.08	18.37	35.59	71.97
BEST-RQ [6]	A	✓	–	–	1.60*	21.90	28.62	77.98
BEST-RQ [6]	A	✓	All	100	5.60	15.32	16.69	68.34
AVFormer (Ours)	**A+V**	✓	VP + Adapters	5	4.36	**13.69**	**16.60**	**64.75**

Results in WER, Lower is better

Google Research

Qualitative Results

Visual information helps to correct ASR mistakes on objects, actions and difficult audio words

Action "slight" corrected to "slice"
Object "carriage" corrected to "carrot"

Object "ball" corrected to "bowl"

Homophone "colonels" corrected to "kernels"

Google Research

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

THU-PM-221

AVFormer: Injecting Vision into Frozen Speech Models for Zero-Shot AV-ASR

Paul Hongsuck Seo, Arsha Nagrani, Cordelia Schmid

Thank You

Google Research