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

Wei-Ning Hsu

Meta AI

Joint work with Bowen Shi, Kushal Lakhotia, Abdelrahman Mohamed

Feb 28, 2022 @ AAAI workshop on Self-Supervised Learning for Audio and Speech Processing

FACEBOOK AI

Motivation

Automatic speech recognition (ASR) is widely used, BUT

- Performance degrades a lot when noisy
- Especially when the noise is speech
- On LRS3, from 4.7% WER (clean) to 32.1% WER (0 dB)

We need to make ASR more robust, how?

- Audio-**visual** speech recognition (AVSR)
  - Use complementary visual information (lip)
  - Invariant to noise
  - In the case of -inf SNR / no audio → lip-reading (VSR)
- VSR is also useful for people with speech impairment

Motivation

Automatic speech recognition (ASR) is widely used, BUT

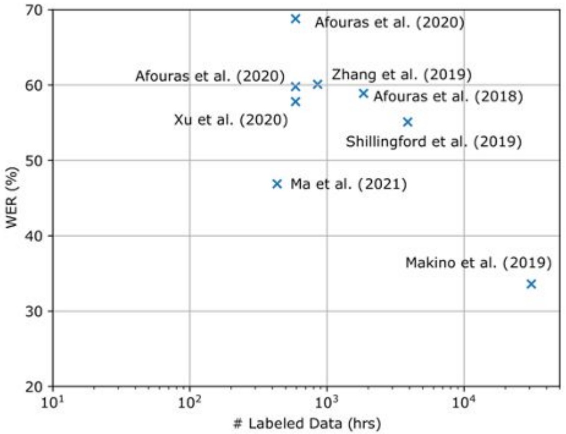
- Performance degrades a lot when noisy
- Especially when the noise is speech
- On LRS3, from 4.7% WER (clean) to 32.1% WER (0 dB)

We need to make ASR more robust, how?

- Audio-**visual** speech recognition (AVSR)
  - Use complementary visual information (lip)
  - Invariant to noise
  - In the case of -inf SNR / no audio → lip-reading (VSR)
- VSR is also useful for people with speech impairment

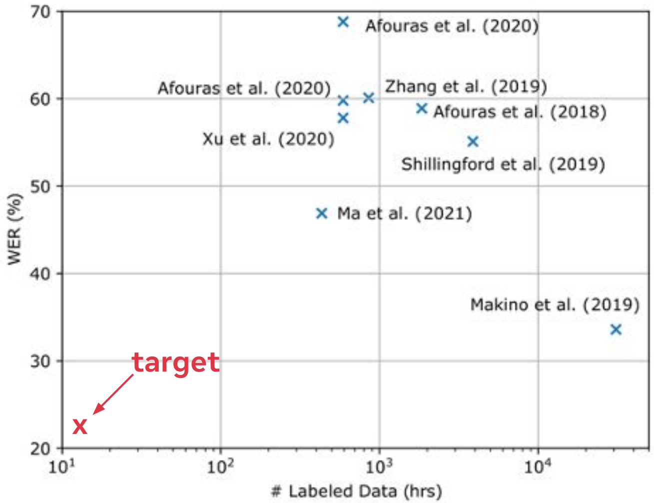
Motivation (cont.)

- What are the barriers?
  - Lack of labeled data
    - Largest public datasets: LRS2* (224), LRS3 (433)
    - Far less than ASR data: Librispeech (1K), GigaSpeech (10K)
    - VSR and AVSR are also data hungry
      - Trained on LRS3: 46.9% WER
      - Trained on YT31k: 33.6% WER
  - Complicated pipeline
    - First pre-trained on isolated words (LRW)
    - Then do curriculum training (short to long)



Motivation (cont.)

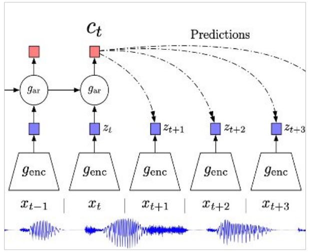
- How do we approach that?
    - Self-supervised learning
        - First, pre-train on unlabeled data
        - Then, fine-tune on (limited) labeled data
    - Are unlabeled data available for VSR/AVSR?
        - Yes! news, movies, social media, meeting, ...
    - What self-supervised learning method to use?
        - Adapt from speech SSL



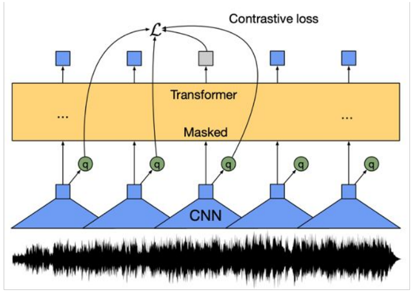
Self-Supervised Learning for Visual Speech Recognition

- What is lip-reading (VSR)?
  - Input: a sequence of image frames
  - output: a sequence of characters/word-pieces
  - Supervised learning: trained on (video, text) pairs

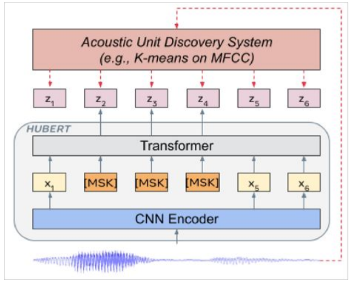
- What speech SSL methods to adapt from?



CPC (Oord et al., 2018)



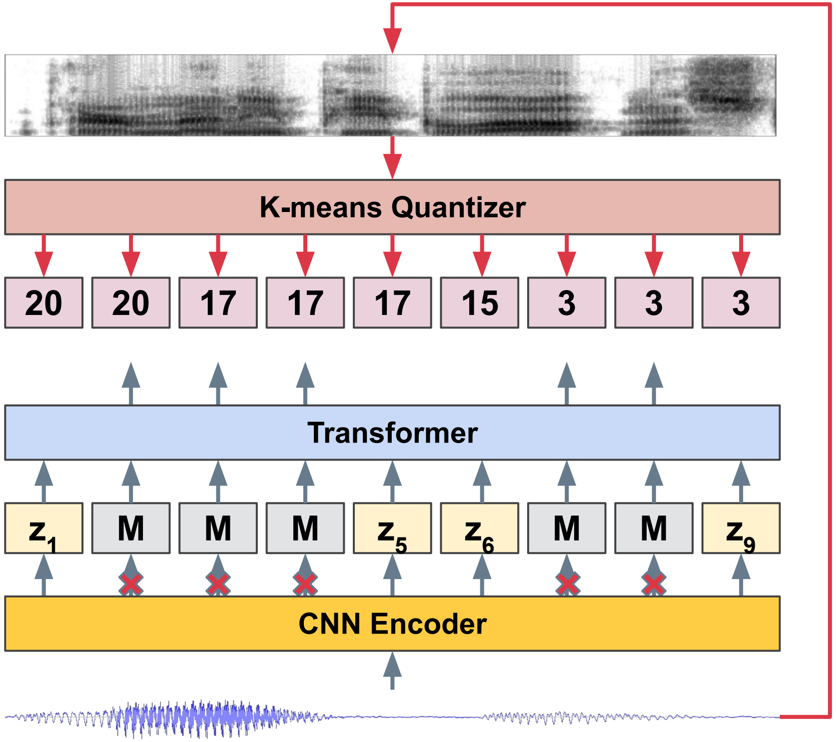
wav2vec 2.0 (Baevski et al., 2021)



HuBERT (Hsu et al., 2021)

...

Preliminary: HuBERT for audio (A-HuBERT)

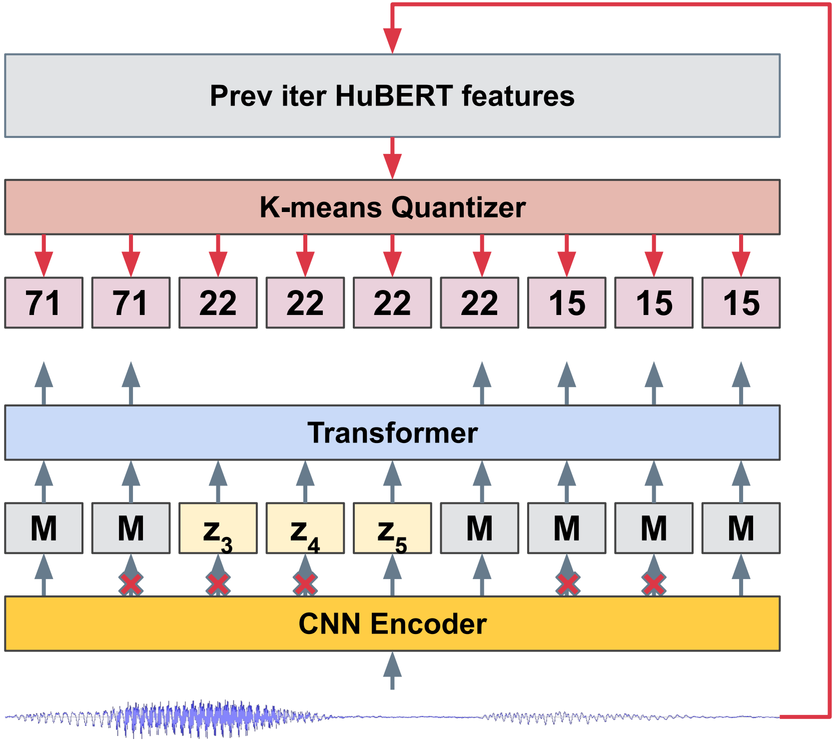


HuBERT is a SOTA SSL framework for speech that performs well on many tasks

How does it work?

- Given an audio stream
  1. K-means on MFCC
  2. Masked prediction

Preliminary: HuBERT for audio (A-HuBERT)

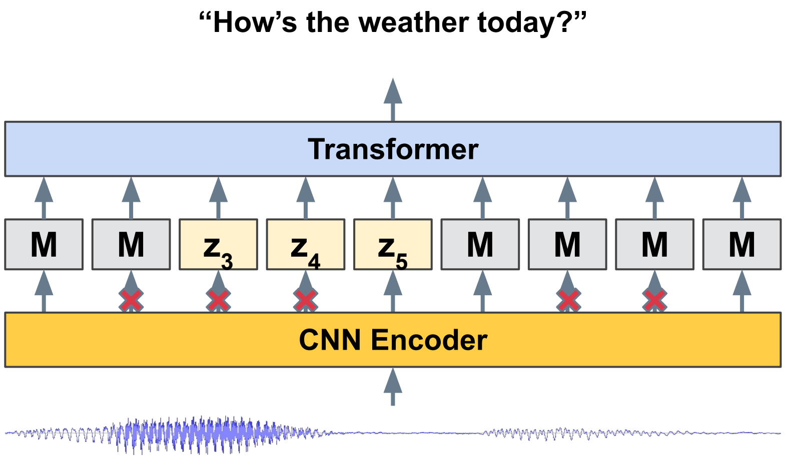


HuBERT is a SOTA SSL framework for speech that performs well on many tasks

How does it work?

- Given an audio stream
  1. K-means on MFCC
  2. Masked prediction
  3. Iterative refinement

Preliminary: HuBERT for audio (A-HuBERT)



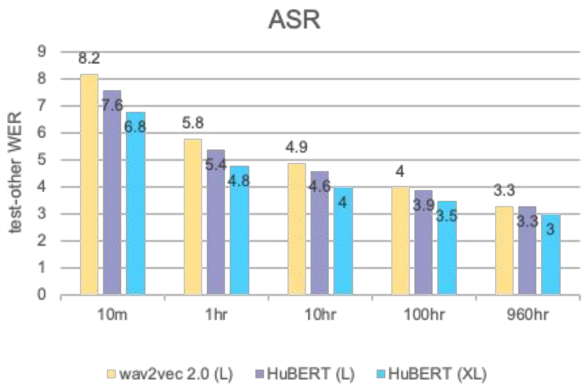
HuBERT is a SOTA SSL framework for speech that performs well on many tasks

How does it work?

- Given an audio stream
    1. K-means on MFCC
    2. Masked prediction
    3. Iterative refinement
    4. Remove cluster prediction head and fine-tune with labeled data

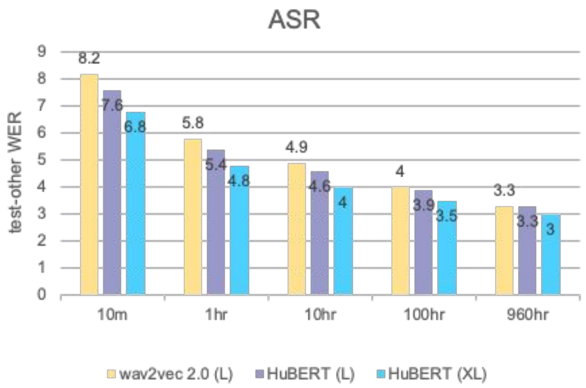
Preliminary: HuBERT for audio (A-HuBERT)

- Shown effective for inference and generative tasks
  - ASR



Preliminary: HuBERT for audio (A-HuBERT)

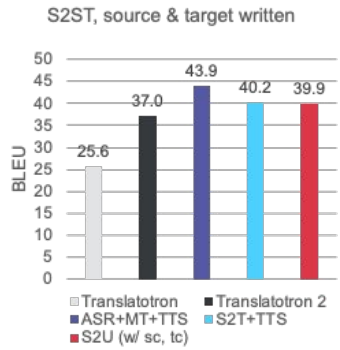
- Shown effective for inference and generative tasks
    - ASR
    - SUPERB (content, speaker, semantic, emotion)

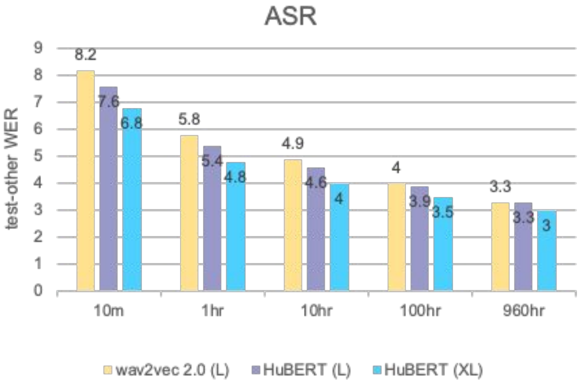


	PR	KS	IC	SID	ER	ASR (WER)		QbE	SF		ASV	SD
	PER ↓	Acc ↑	Acc ↑	Acc ↑	Acc ↑	w/o ↓	w/ LM ↓	MTWV ↑	F1 ↑	CER ↓	EER ↓	DER ↓
FBANK	82.01	8.63	9.10	8.5E-4	35.39	23.18	15.21	0.0058	69.64	52.94	9.56	10.05
PASE+ [16]	58.87	82.54	29.82	37.99	57.86	25.11	16.62	0.0072	62.14	60.17	11.61	8.68
APC [7]	41.98	91.01	74.69	60.42	59.33	21.28	14.74	0.0310	70.46	50.89	8.56	10.53
VQ-APC [32]	41.08	91.11	74.48	60.15	59.66	21.20	15.21	0.0251	68.53	52.91	8.72	10.45
NPC [33]	43.81	88.96	69.44	55.92	59.08	20.20	13.91	0.0246	72.79	48.44	9.4	9.34
Mockingjay [8]	70.19	83.67	34.33	32.29	50.28	22.82	15.48	6.6E-04	61.59	58.89	11.66	10.54
TERA [9]	49.17	89.48	58.42	57.57	56.27	18.17	12.16	0.0013	67.50	54.17	15.89	9.96
modified CPC [34]	42.54	91.88	64.09	39.63	60.96	20.18	13.53	0.0326	71.19	49.91	12.86	10.38
wav2vec [12]	31.58	95.59	84.92	56.56	59.79	15.86	11.00	0.0485	76.37	43.71	7.99	9.9
vq-wav2vec [13]	33.48	93.38	85.68	38.80	58.24	17.71	12.80	0.0410	77.68	41.54	10.38	9.93
wav2vec 2.0 Base [14]	5.74	96.23	92.35	75.18	63.43	6.43	4.79	0.0233	88.30	24.77	6.02	6.08
wav2vec 2.0 Large [14]	4.75	**96.66**	95.28	86.14	65.64	3.75	3.10	0.0489	87.11	27.31	5.65	**5.62**
HuBERT Base [35]	5.41	96.30	98.34	81.42	64.92	6.42	4.79	**0.0736**	88.53	25.20	**5.11**	5.88
HuBERT Large [35]	**3.53**	95.29	**98.76**	**90.33**	**67.62**	**3.62**	**2.94**	0.0353	**89.81**	**21.76**	5.98	5.75

Preliminary: HuBERT for audio (A-HuBERT)

- Shown effective for inference and generative tasks
  - ASR
  - SUPERB (content, speaker, semantic, emotion)
  - Speech-to-speech translation

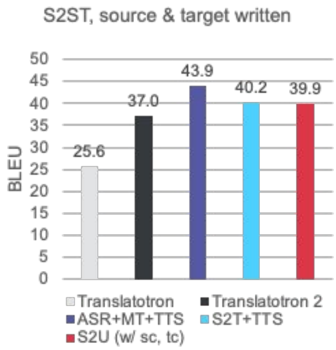




	PR	KS	IC	SID	ER	ASR (WER)		QbE	SF		ASV	SD
	PER ↓	Acc ↑	Acc ↑	Acc ↑	Acc ↑	w/o ↓	w/ LM ↓	MTWV ↑	F1 ↑	CER ↓	EER ↓	DER ↓
FBANK	82.01	8.63	9.10	8.5E-4	35.39	23.18	15.21	0.0058	69.64	52.94	9.56	10.05
PASE+ [16]	58.87	82.54	29.82	37.99	57.86	25.11	16.62	0.0072	62.14	60.17	11.61	8.68
APC [7]	41.98	91.01	74.69	60.42	59.33	21.28	14.74	0.0310	70.46	50.89	8.56	10.53
VQ-APC [32]	41.08	91.11	74.48	60.15	59.66	21.20	15.21	0.0251	68.53	52.91	8.72	10.45
NPC [33]	43.81	88.96	69.44	55.92	59.08	20.20	13.91	0.0246	72.79	48.44	9.4	9.34
Mockingjay [8]	70.19	83.67	34.33	32.29	50.28	22.82	15.48	6.6E-04	61.59	58.89	11.66	10.54
TERA [9]	49.17	89.48	58.42	57.57	56.27	18.17	12.16	0.0013	67.50	54.17	15.89	9.96
modified CPC [34]	42.54	91.88	64.09	39.63	60.96	20.18	13.53	0.0326	71.19	49.91	12.86	10.38
wav2vec [12]	31.58	95.59	84.92	56.56	59.79	15.86	11.00	0.0485	76.37	43.71	7.99	9.9
vq-wav2vec [13]	33.48	93.38	85.68	38.80	58.24	17.71	12.80	0.0410	77.68	41.54	10.38	9.93
wav2vec 2.0 Base [14]	5.74	96.23	92.35	75.18	63.43	6.43	4.79	0.0233	88.30	24.77	6.02	6.08
wav2vec 2.0 Large [14]	4.75	**96.66**	95.28	86.14	65.64	3.75	3.10	0.0489	87.11	27.31	5.65	**5.62**
HuBERT Base [35]	5.41	96.30	98.34	81.42	64.92	6.42	4.79	**0.0736**	88.53	25.20	**5.11**	5.88
HuBERT Large [35]	**3.53**	95.29	**98.76**	**90.33**	**67.62**	**3.62**	**2.94**	0.0353	**89.81**	**21.76**	5.98	5.75

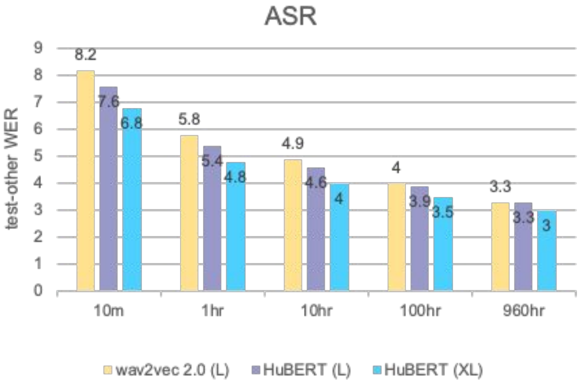
Preliminary: HuBERT for audio (A-HuBERT)

- Shown effective for inference and generative tasks
  - ASR
  - SUPERB (content, speaker, semantic, emotion)
  - Speech-to-speech translation
  - Textless NLP (https://speechbot.github.io):
    - text-free audio GPT



Prompt

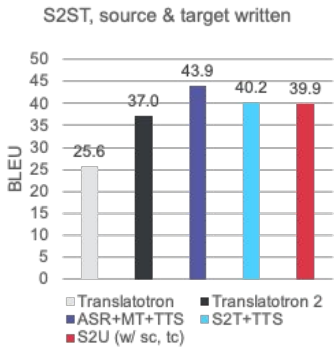
Generated

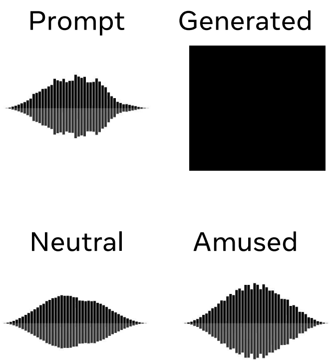


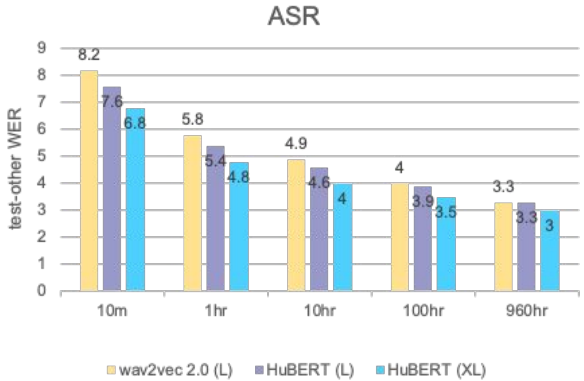
	PR	KS	IC	SID	ER	ASR (WER)		QbE	SF		ASV	SD
	PER ↓	Acc ↑	Acc ↑	Acc ↑	Acc ↑	w/o ↓	w/ LM ↓	MTWV ↑	F1 ↑	CER ↓	EER ↓	DER ↓
FBANK	82.01	8.63	9.10	8.5E-4	35.39	23.18	15.21	0.0058	69.64	52.94	9.56	10.05
PASE+ [16]	58.87	82.54	29.82	37.99	57.86	25.11	16.62	0.0072	62.14	60.17	11.61	8.68
APC [7]	41.98	91.01	74.69	60.42	59.33	21.28	14.74	0.0310	70.46	50.89	8.56	10.53
VQ-APC [32]	41.08	91.11	74.48	60.15	59.66	21.20	15.21	0.0251	68.53	52.91	8.72	10.45
NPC [33]	43.81	88.96	69.44	55.92	59.08	20.20	13.91	0.0246	72.79	48.44	9.4	9.34
Mockingjay [8]	70.19	83.67	34.33	32.29	50.28	22.82	15.48	6.6E-04	61.59	58.89	11.66	10.54
TERA [9]	49.17	89.48	58.42	57.57	56.27	18.17	12.16	0.0013	67.50	54.17	15.89	9.96
modified CPC [34]	42.54	91.88	64.09	39.63	60.96	20.18	13.53	0.0326	71.19	49.91	12.86	10.38
wav2vec [12]	31.58	95.59	84.92	56.56	59.79	15.86	11.00	0.0485	76.37	43.71	7.99	9.9
vq-wav2vec [13]	33.48	93.38	85.68	38.80	58.24	17.71	12.80	0.0410	77.68	41.54	10.38	9.93
wav2vec 2.0 Base [14]	5.74	96.23	92.35	75.18	63.43	6.43	4.79	0.0233	88.30	24.77	6.02	6.08
wav2vec 2.0 Large [14]	4.75	**96.66**	95.28	86.14	65.64	3.75	3.10	0.0489	87.11	27.31	5.65	**5.62**
HuBERT Base [35]	5.41	96.30	98.34	81.42	64.92	6.42	4.79	**0.0736**	88.53	25.20	**5.11**	5.88
HuBERT Large [35]	**3.53**	95.29	**98.76**	**90.33**	**67.62**	**3.62**	**2.94**	0.0353	**89.81**	**21.76**	5.98	5.75

Preliminary: HuBERT for audio (A-HuBERT)

- Shown effective for inference and generative tasks
  - ASR
  - SUPERB (content, speaker, semantic, emotion)
  - Speech-to-speech translation
  - Textless NLP (https://speechbot.github.io):
    - text-free audio GPT, emotion conversion

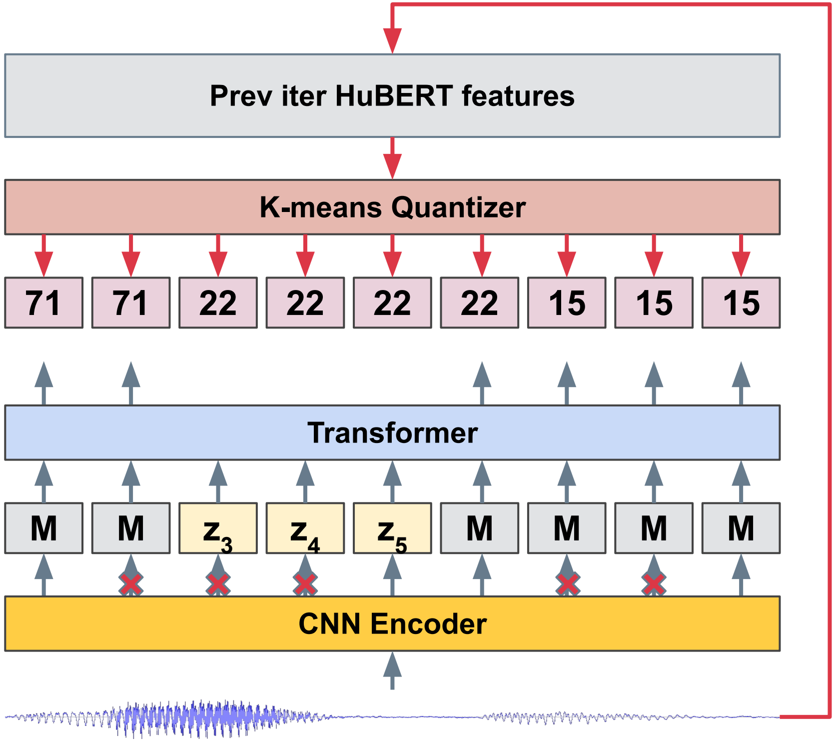




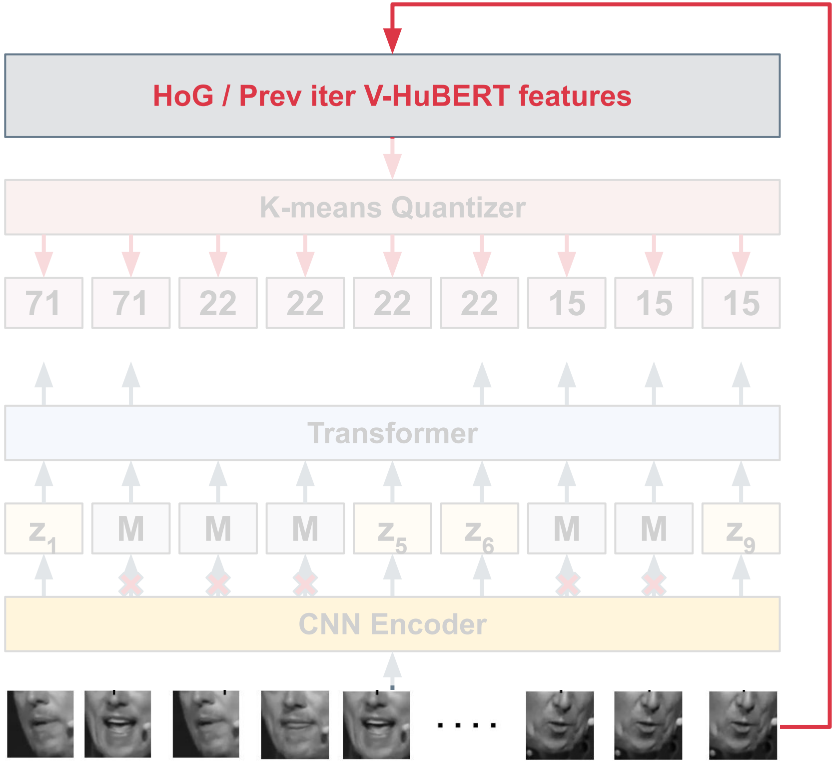


	PR	KS	IC	SID	ER	ASR (WER)		QbE	SF		ASV	SD
	PER ↓	Acc ↑	Acc ↑	Acc ↑	Acc ↑	w/o ↓	w/ LM ↓	MTWV ↑	F1 ↑	CER ↓	EER ↓	DER ↓
FBANK	82.01	8.63	9.10	8.5E-4	35.39	23.18	15.21	0.0058	69.64	52.94	9.56	10.05
PASE+ [16]	58.87	82.54	29.82	37.99	57.86	25.11	16.62	0.0072	62.14	60.17	11.61	8.68
APC [7]	41.98	91.01	74.69	60.42	59.33	21.28	14.74	0.0310	70.46	50.89	8.56	10.53
VQ-APC [32]	41.08	91.11	74.48	60.15	59.66	21.20	15.21	0.0251	68.53	52.91	8.72	10.45
NPC [33]	43.81	88.96	69.44	55.92	59.08	20.20	13.91	0.0246	72.79	48.44	9.4	9.34
Mockingjay [8]	70.19	83.67	34.33	32.29	50.28	22.82	15.48	6.6E-04	61.59	58.89	11.66	10.54
TERA [9]	49.17	89.48	58.42	57.57	56.27	18.17	12.16	0.0013	67.50	54.17	15.89	9.96
modified CPC [34]	42.54	91.88	64.09	39.63	60.96	20.18	13.53	0.0326	71.19	49.91	12.86	10.38
wav2vec [12]	31.58	95.59	84.92	56.56	59.79	15.86	11.00	0.0485	76.37	43.71	7.99	9.9
vq-wav2vec [13]	33.48	93.38	85.68	38.80	58.24	17.71	12.80	0.0410	77.68	41.54	10.38	9.93
wav2vec 2.0 Base [14]	5.74	96.23	92.35	75.18	63.43	6.43	4.79	0.0233	88.30	24.77	6.02	6.08
wav2vec 2.0 Large [14]	4.75	**96.66**	95.28	86.14	65.64	3.75	3.10	0.0489	87.11	27.31	5.65	**5.62**
HuBERT Base [35]	5.41	96.30	98.34	81.42	64.92	6.42	4.79	**0.0736**	88.53	25.20	**5.11**	5.88
HuBERT Large [35]	**3.53**	95.29	**98.76**	**90.33**	**67.62**	**3.62**	**2.94**	0.0353	**89.81**	**21.76**	5.98	5.75

Preliminary: HuBERT for audio (A-HuBERT)

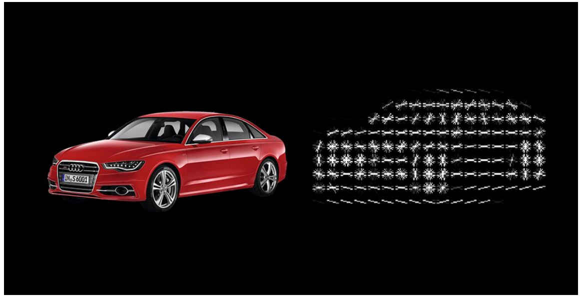


Single-modal Visual HuBERT



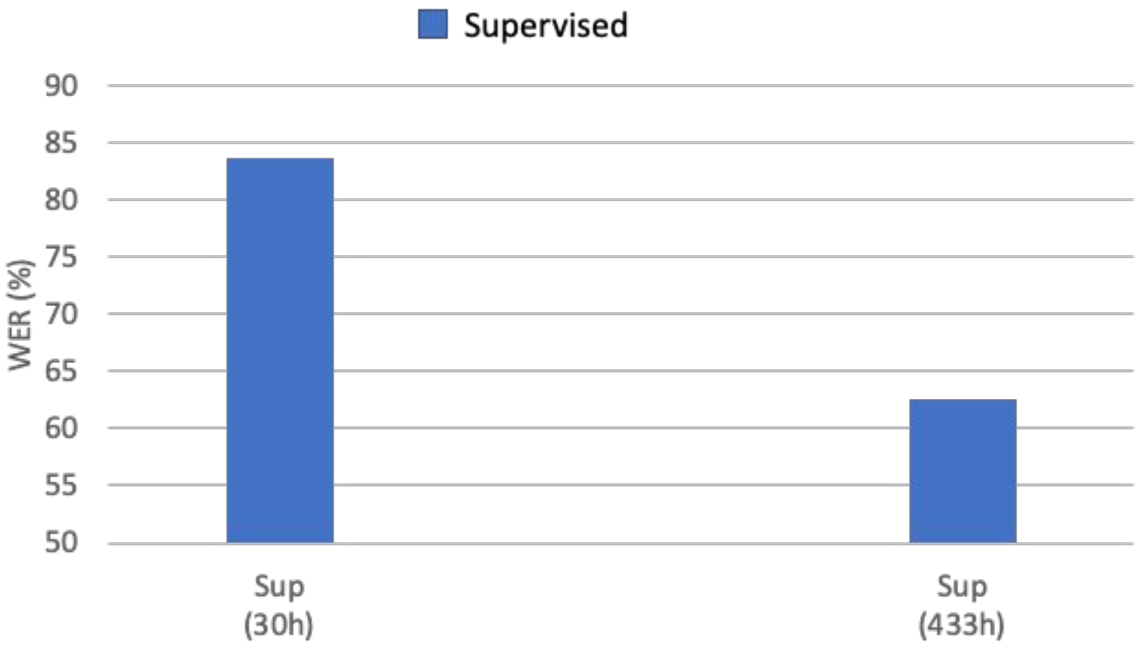
Our initial attempt, given a video stream

1. Use ~~waveform~~ images as input
2. Cluster ~~MFCC~~ Histogram of Gradient (HoG) features



HoG example (source: https://www.analyticsvidhya.com/blog/2019/09/feature-engineering-images-introduction-hog-feature-descriptor)

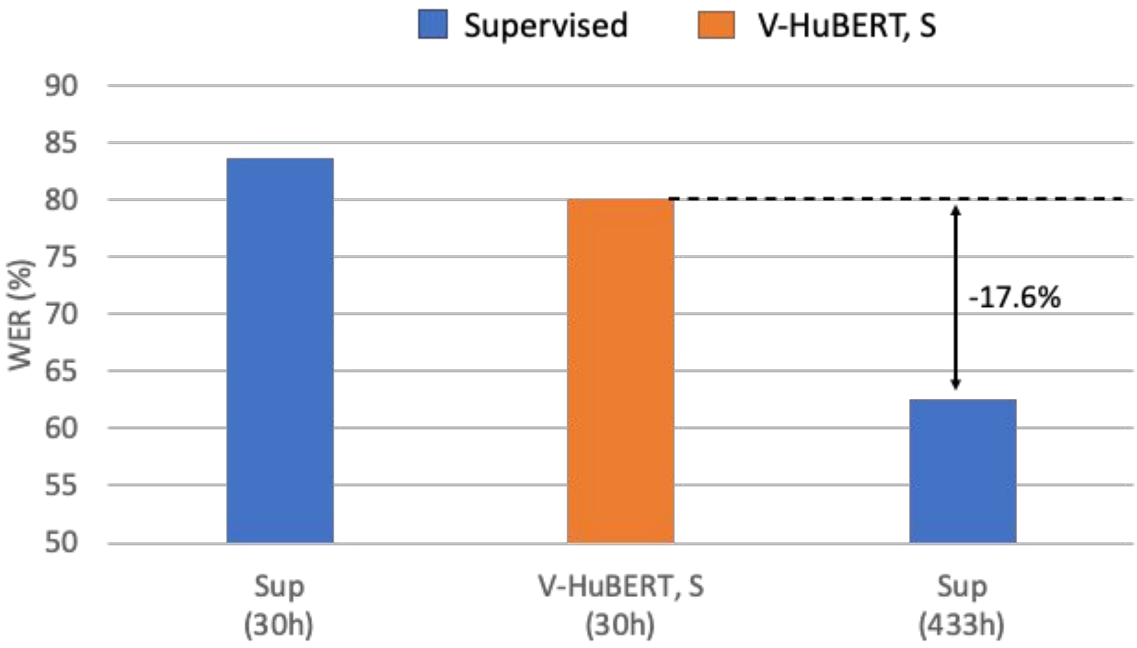
Single-modal Visual HuBERT



- Pre-train:
  - LRS3-433h unlabeled
- Fine-tune:
  - LRS3-30h labeled
  - CTC
- Arch: BERT Base (12L)

- Supervised baselines
  - 30h: 84%
  - 433h: 62.5%

Single-modal Visual HuBERT

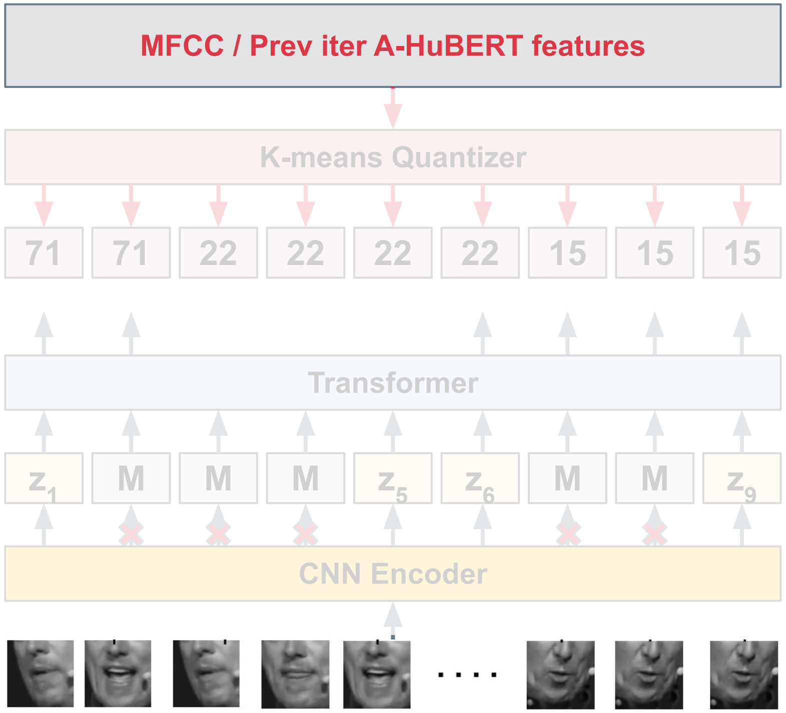


- Pre-train:
  - LRS3-433h unlabeled
- Fine-tune:
  - LRS3-30h labeled
  - CTC
- Arch: BERT Base (12L)

Limited improvement (80.1%):

- HoG cluster quality is “bad”
- Improve it with audio?

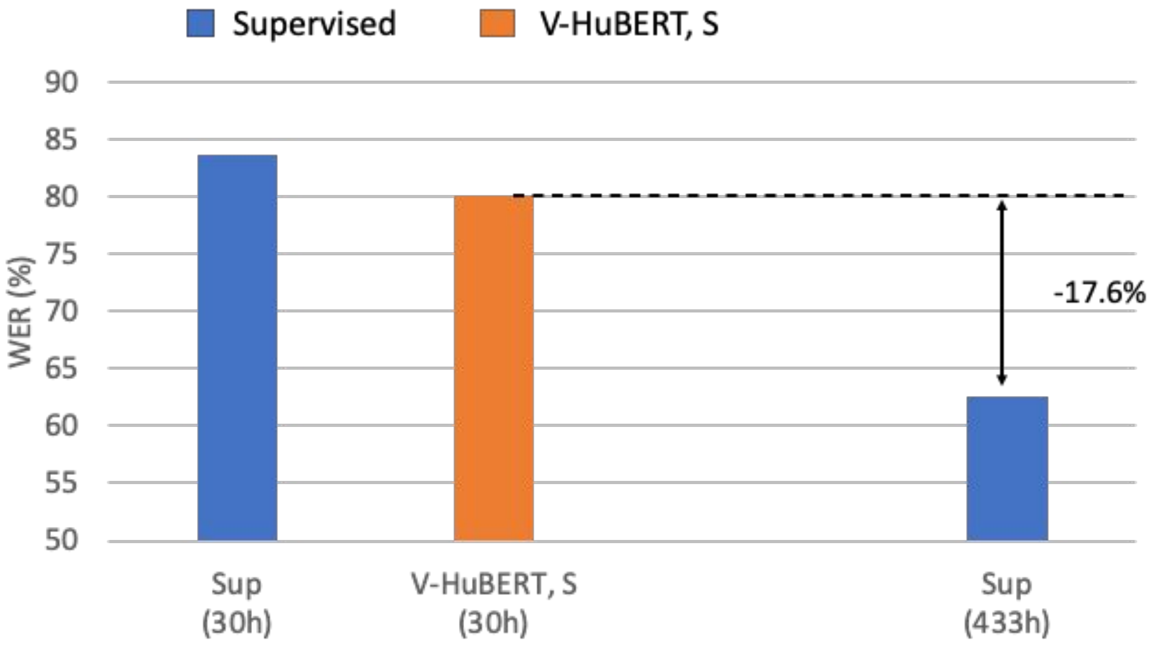
Cross-modal Visual HuBERT



Our second attempt, given audio-visual streams

1. Use lip ROI as input
2. Cluster ~~HoG~~ MFCC/A-HuBERT features
   a. Leverage frame synchronicity to determine the V-A alignment

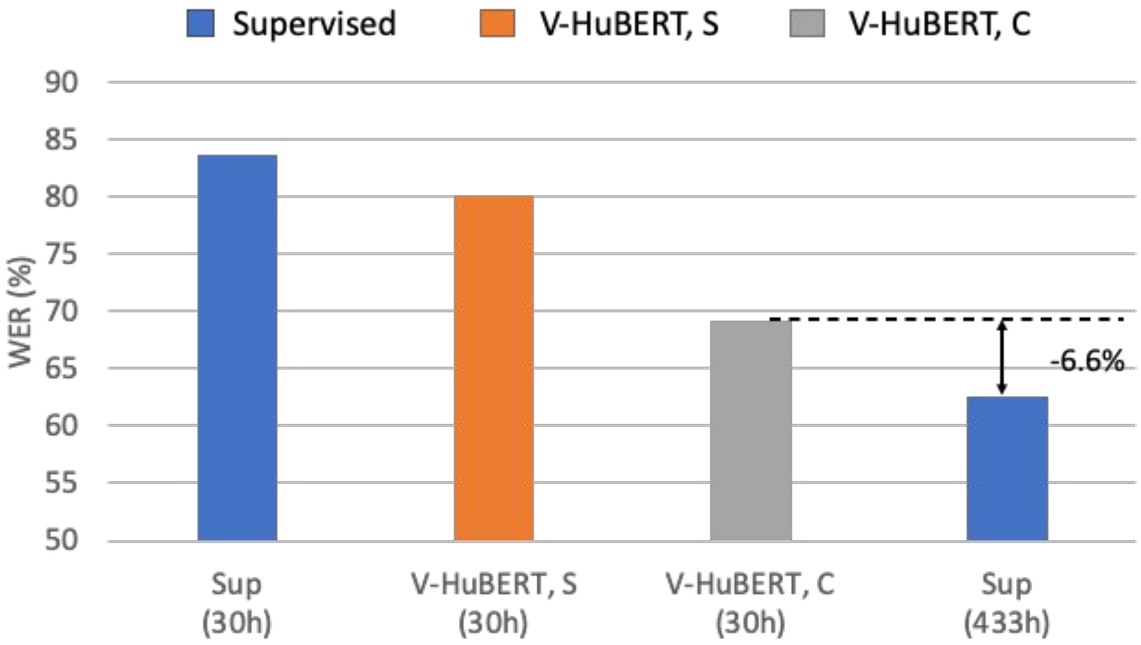
Cross-modal Visual HuBERT



Baselines:

- Sup 30h: 83.7%
- V-HuBERT, S 30h: 80.1%
- Sup 433h: 62.5%

Cross-modal Visual HuBERT

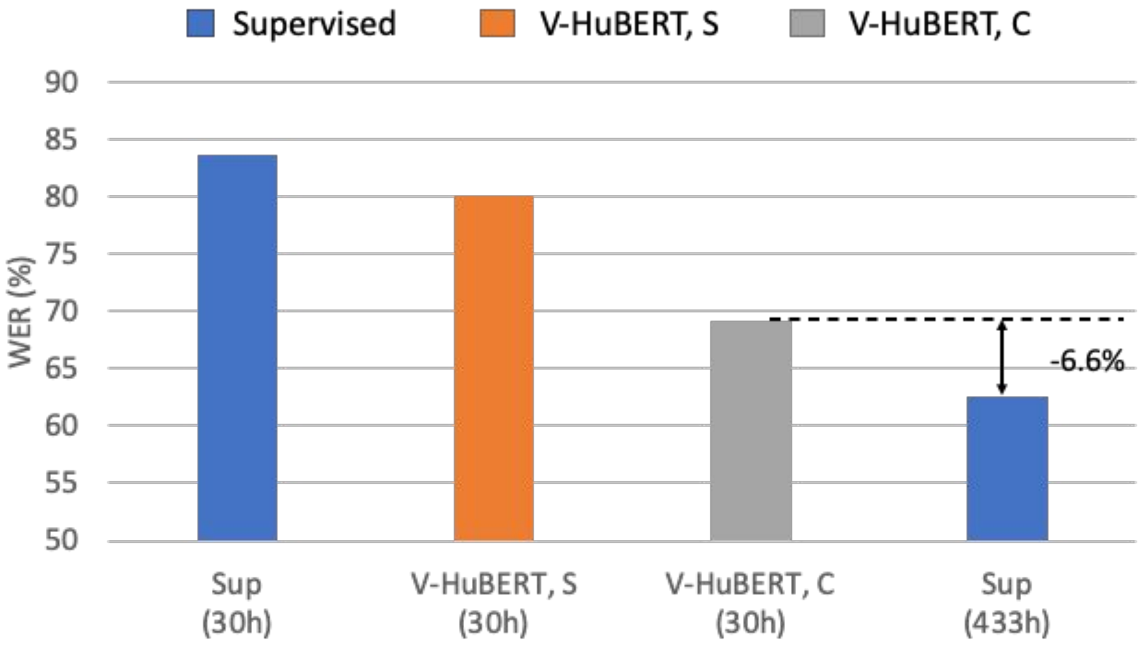


Baselines:

- Sup 30h: 83.7%
- V-HuBERT, S 30h: 80.1%
- V-HuBERT, C 30h: 69.1%
- Sup 433h: 62.5%

→ Predicting audio clusters improves visual representation learning

Cross-modal Visual HuBERT



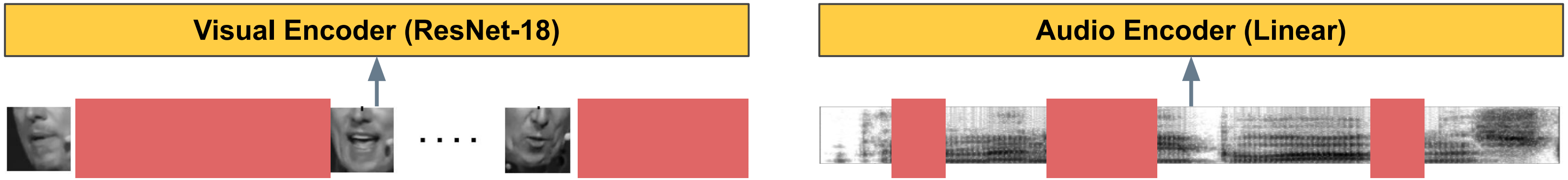
audio cluster > video cluster

What about audio-visual cluster?

- Need AV-features
- Train HuBERT with A+V input
- Can we still fine-tune it for VSR?

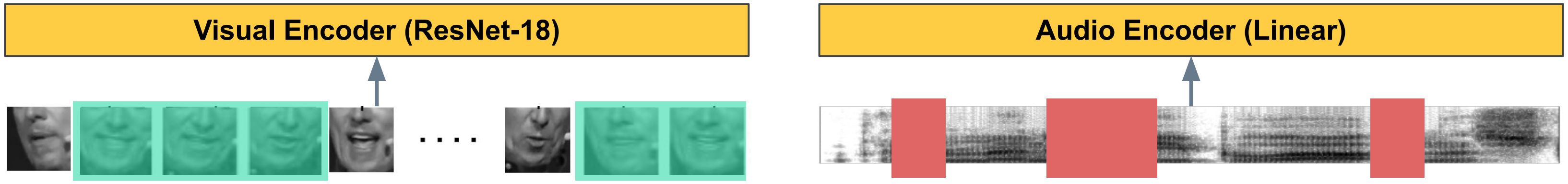
Proposed: Audio-Visual HuBERT

- Use both audio and visual streams input
    - Mask at input independently



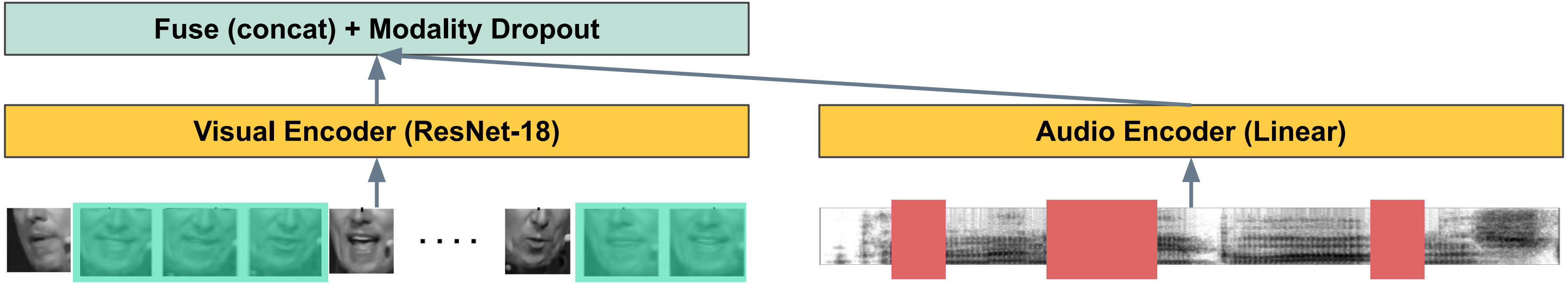
Proposed: Audio-Visual HuBERT

- Use both audio and visual streams input
  - Mask at input independently
  - Mask visual segments by substitution

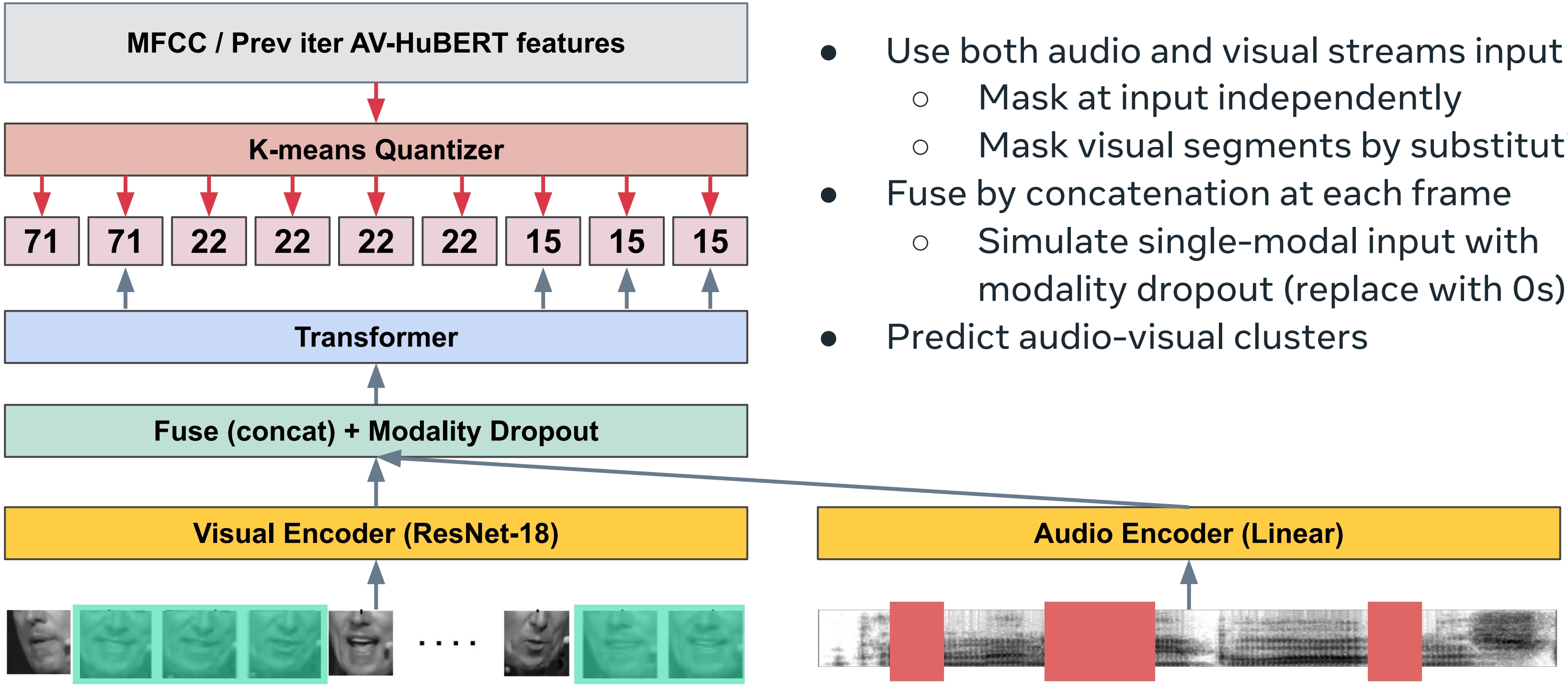


Proposed: Audio-Visual HuBERT

- Use both audio and visual streams input
  - Mask at input independently
  - Mask visual segments by substitution
- Fuse by concatenation at each frame
  - Simulate single-modal input with modality dropout (replace with 0s)

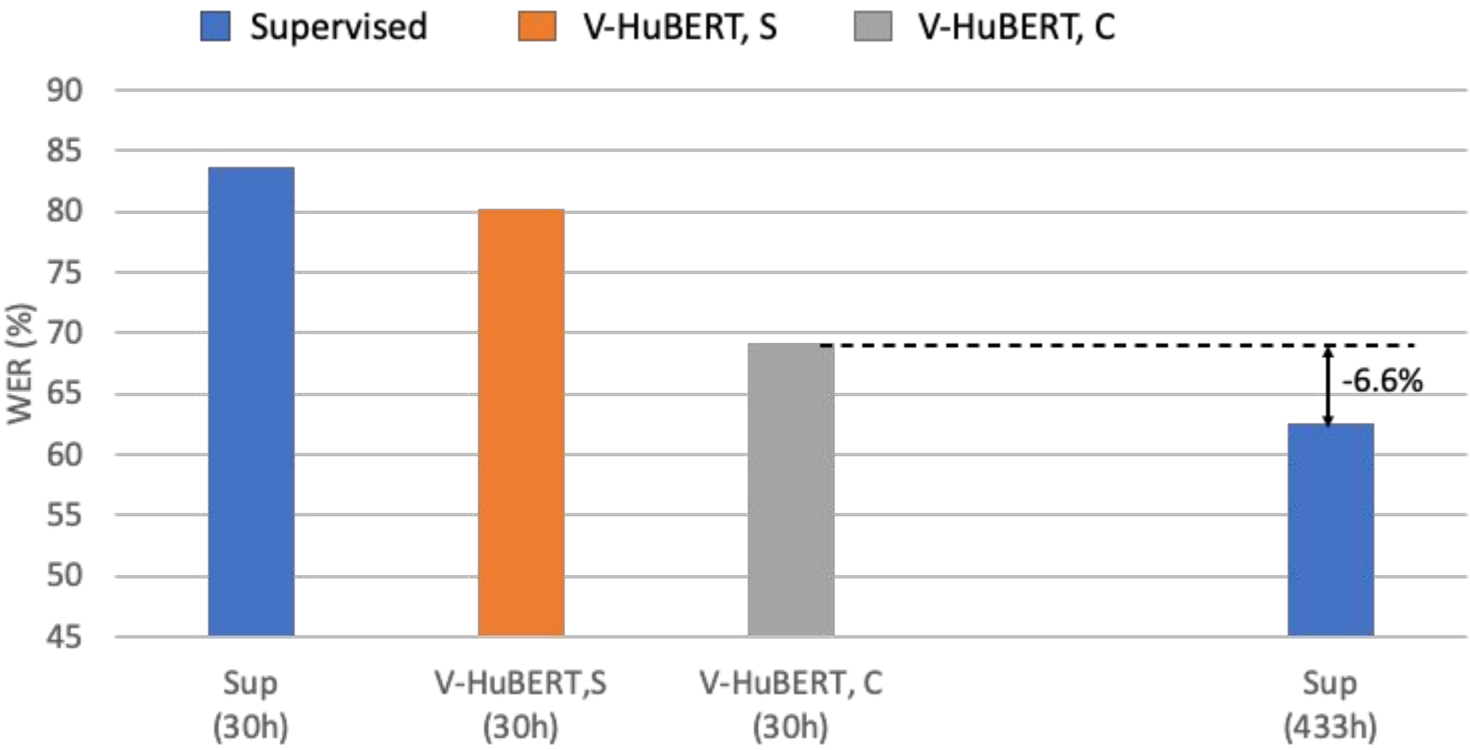


Proposed: Audio-Visual HuBERT

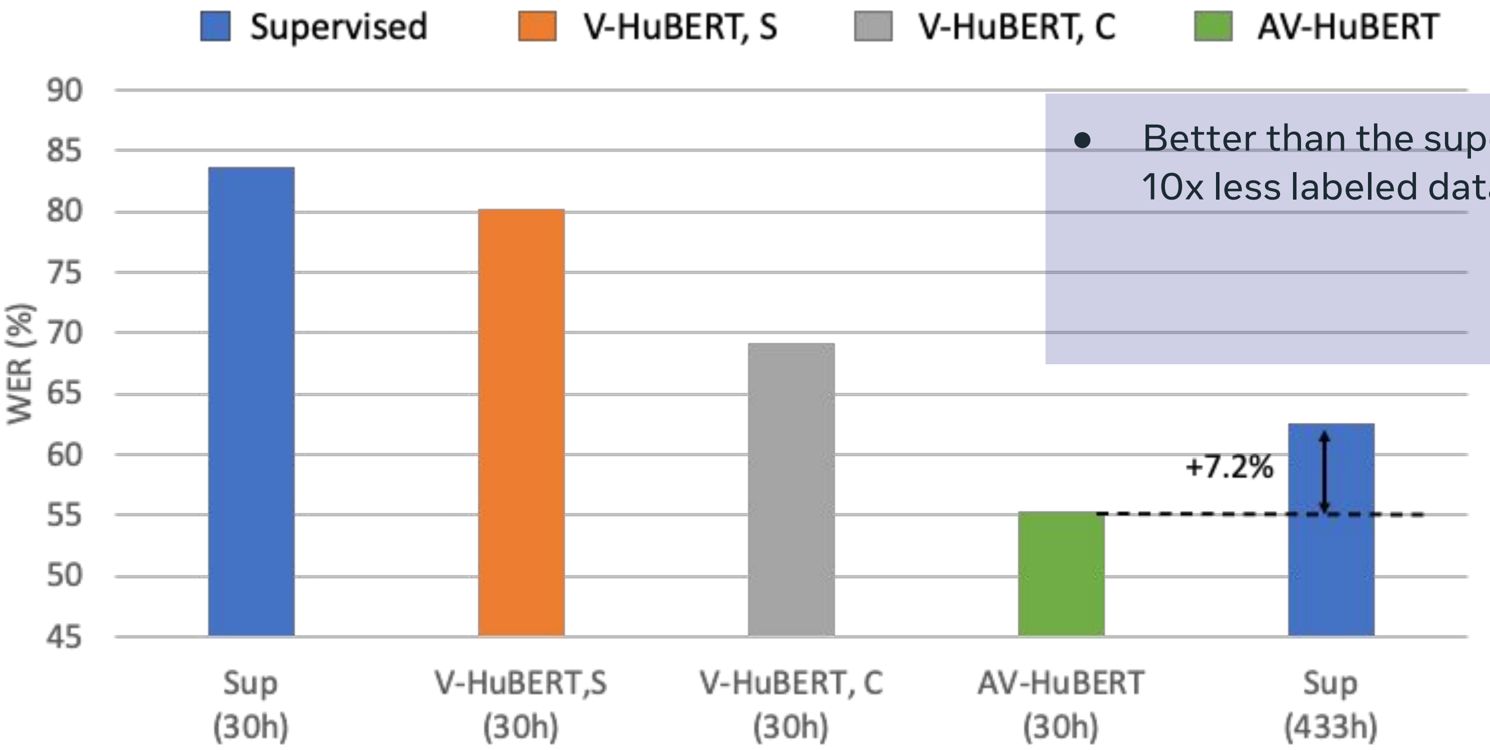


- Use both audio and visual streams input
  - Mask at input independently
  - Mask visual segments by substitution
- Fuse by concatenation at each frame
  - Simulate single-modal input with modality dropout (replace with 0s)
- Predict audio-visual clusters

Proposed: Audio-Visual HuBERT

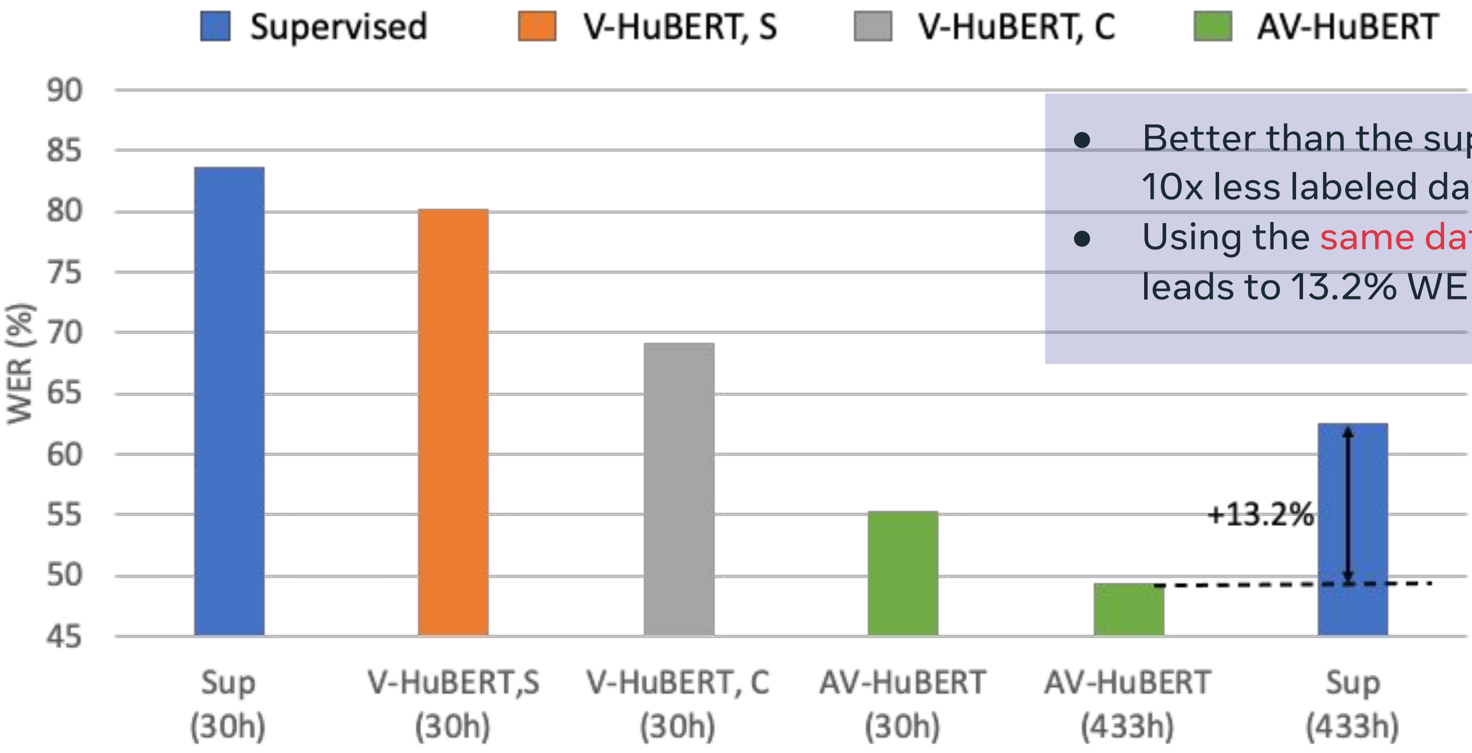


Proposed: Audio-Visual HuBERT



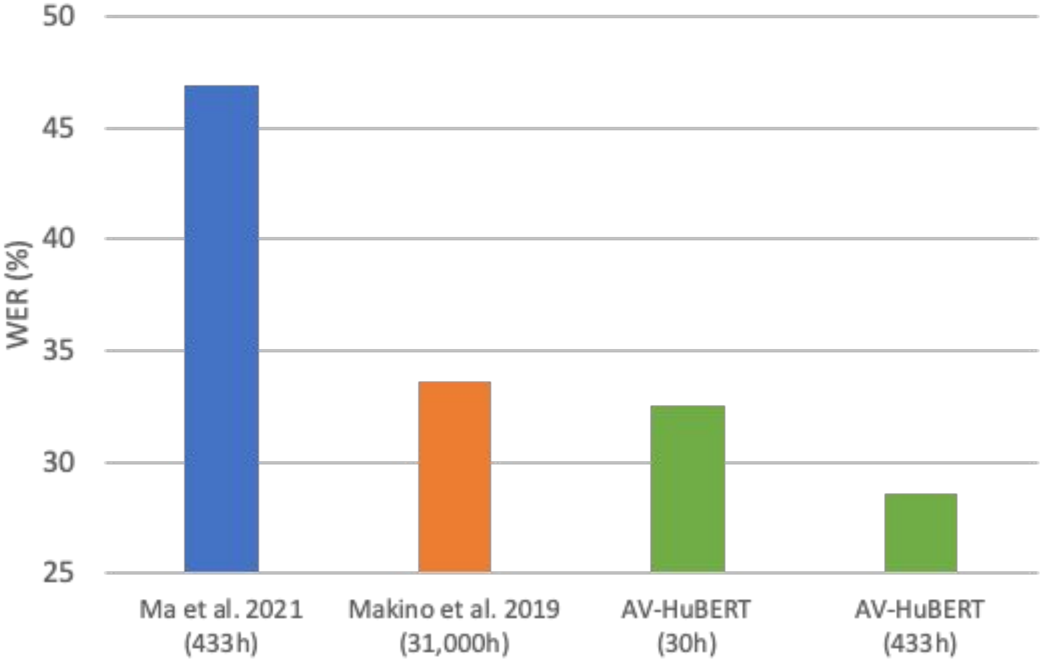
- Better than the supervised model with 10x less labeled data (55.3%)

Proposed: Audio-Visual HuBERT



- Better than the supervised model with 10x less labeled data (55.3%)
- Using the same data*, pre-training leads to 13.2% WER reduction (49.3%)

Comparison with prior works



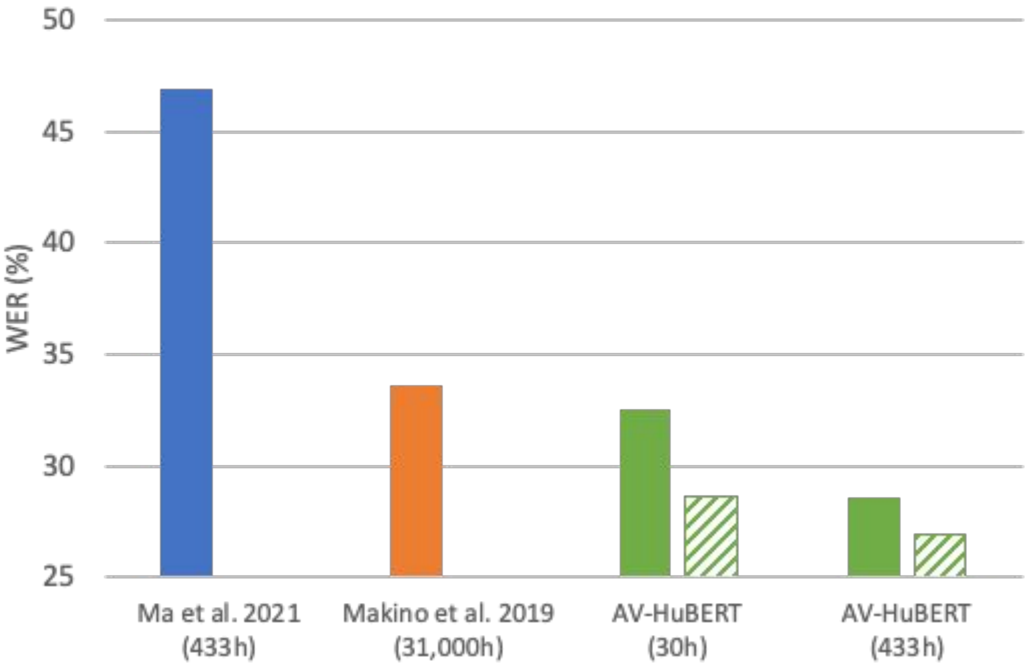
Scaling up

- More unlabeled data (1.7k hours, LRS3 + VoxCeleb2)
- Bigger model (24L BERT LARGE)
- Seq2seq fine-tuning (9L decoder)

Results:

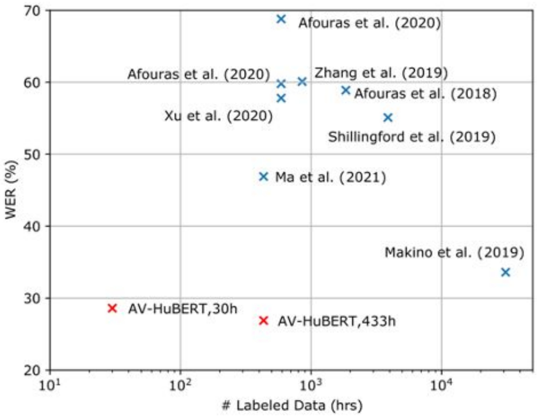
- 30h SSL > 31K hours supervised
- Further improvement with 433hr

Comparison with prior works



Complimentary to self-training

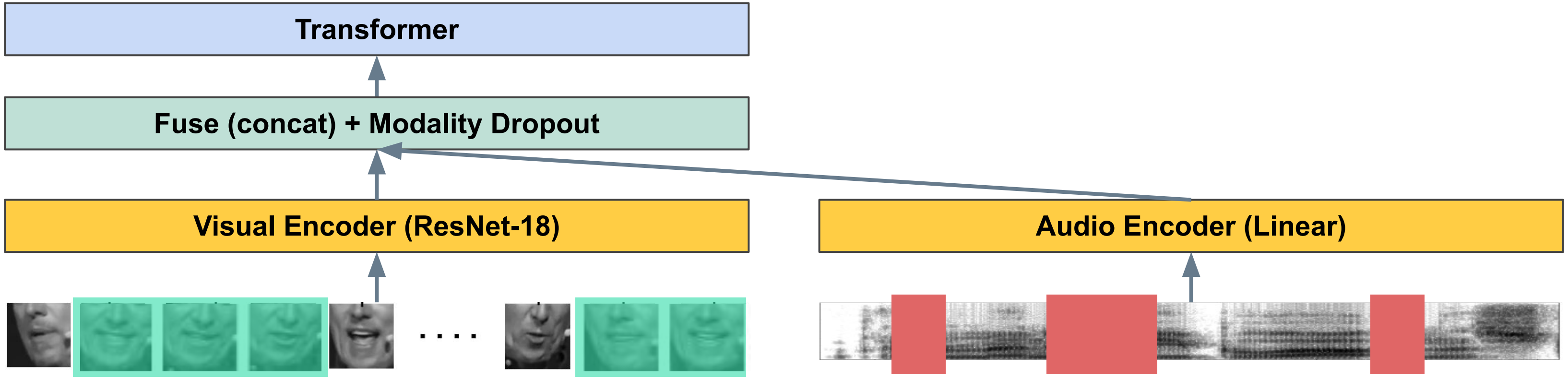
- With 30h, 32.5% -> 28.6%
- With 433h, 28.6% -> 26.9%
  - New SOTA



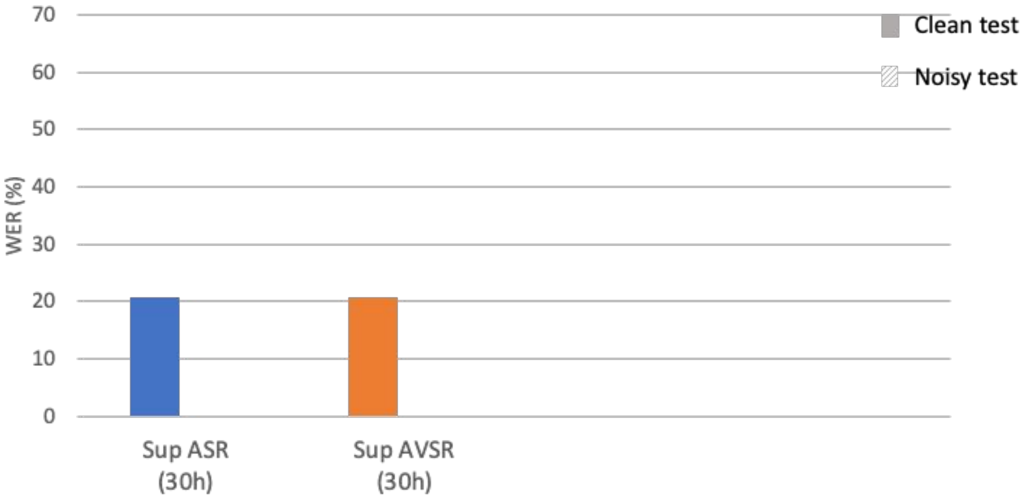
How Effective AV-HuBERT is for Audio-Visual Speech Recognition?

- Audio-visual speech recognition (AVSR)
  - Input: audio+video streams
  - output: a sequence of characters/word-pieces
  - Supervised learning: trained on (audio, video, text) tuples

- AV-HuBERT is a natural fit for audio-visual speech recognition



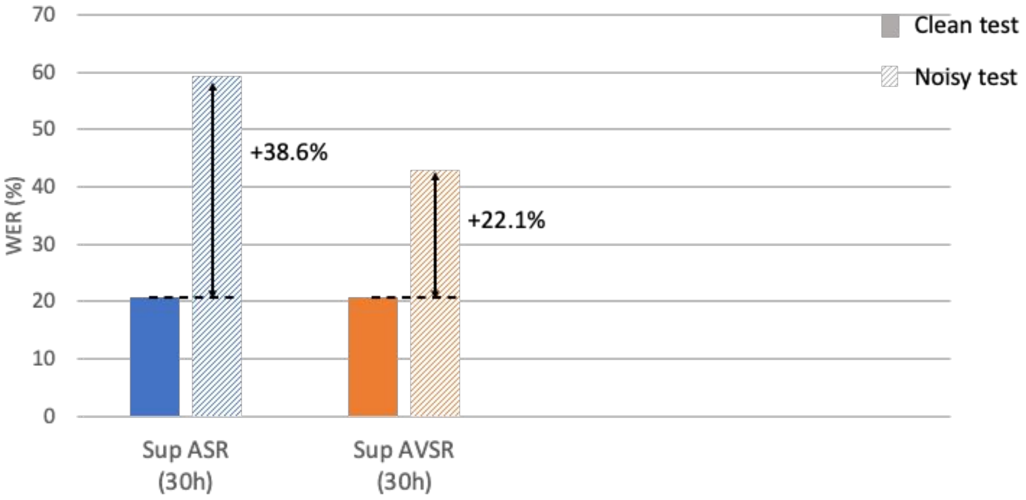
Supervised baselines



- BERT Large (24L enc + 9L dec)
- Supervised fine-tune:
  - LRS3-30h labeled
  - Seq2Seq

- When tested on clean data:
  - ASR: 20.6% WER
  - AVSR: 20.8% WER
- Similar performance in clean cond.

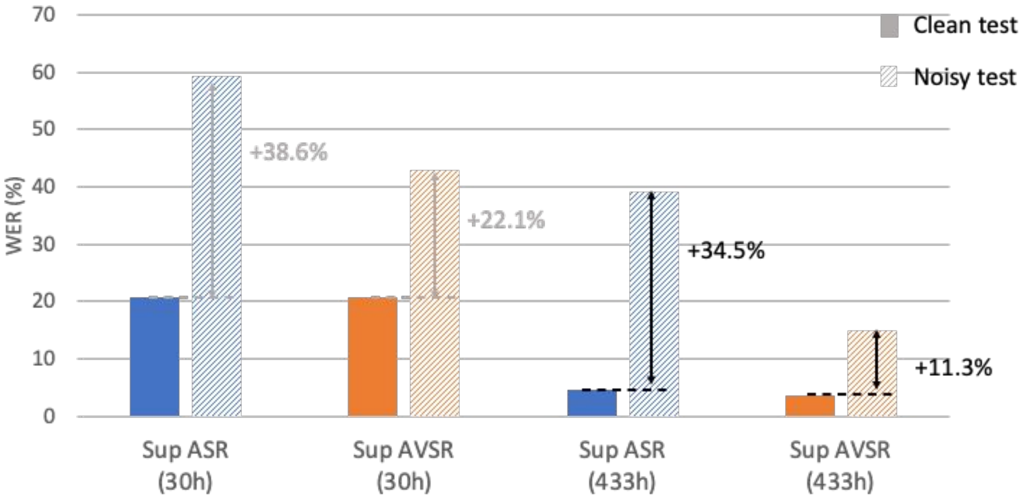
Supervised baselines



- When tested on noisy data:
  - ASR: +38.6% WER
  - AVSR: +22.1% WER

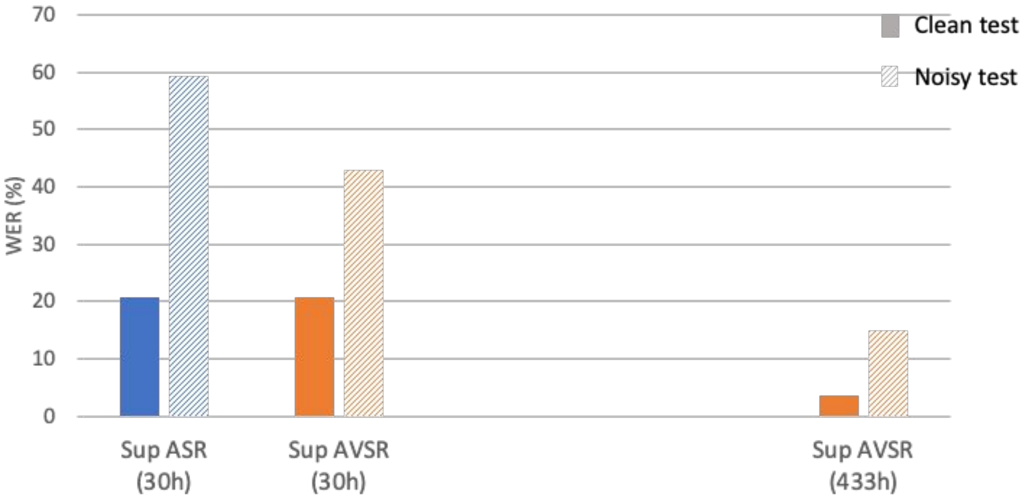
Both degrades, but AVSR is more robust

Supervised baselines

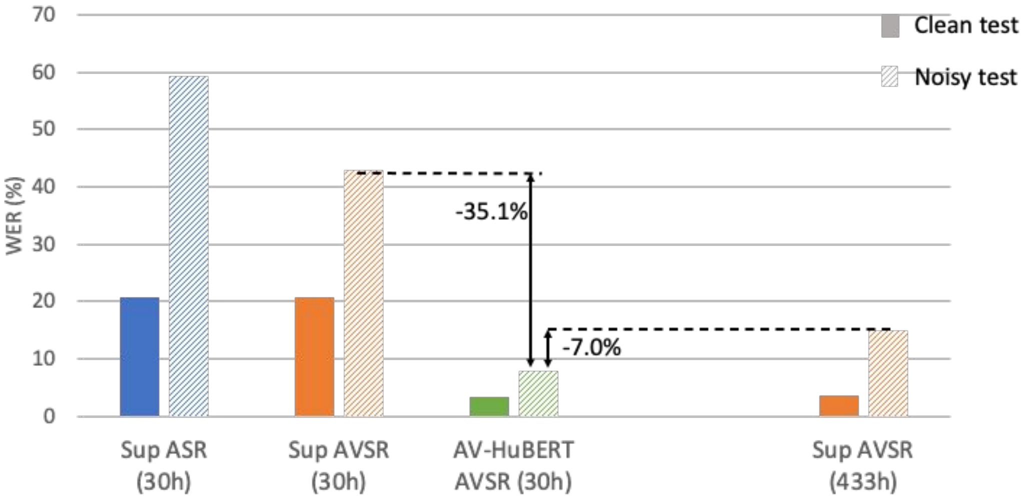


- Same trend when increasing data

AV-HuBERT results of AVSR

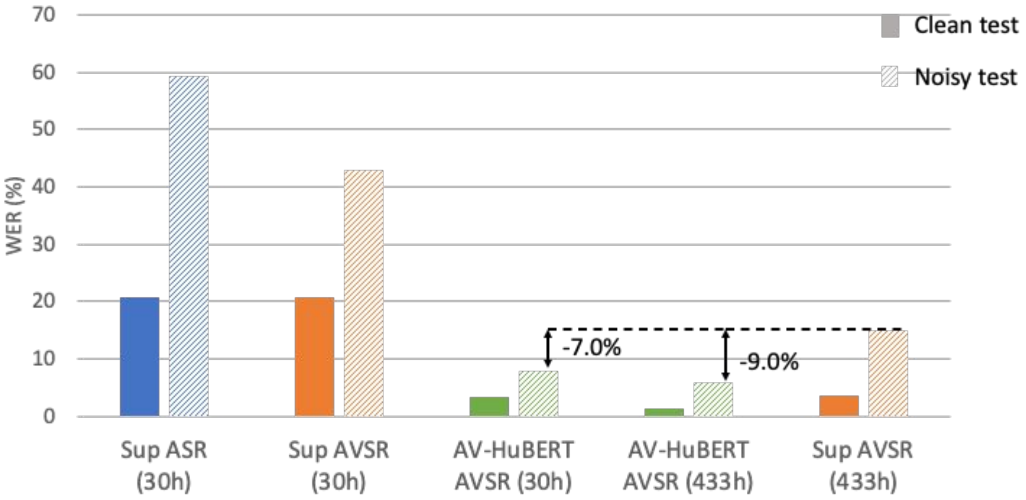


AV-HuBERT results of AVSR



- 35.1% absolute WER reduction
- beats 433hr-supervised model on with just 30hr labeled
- More robust to noise
  - 3.3% on clean, 7.8% on noisy

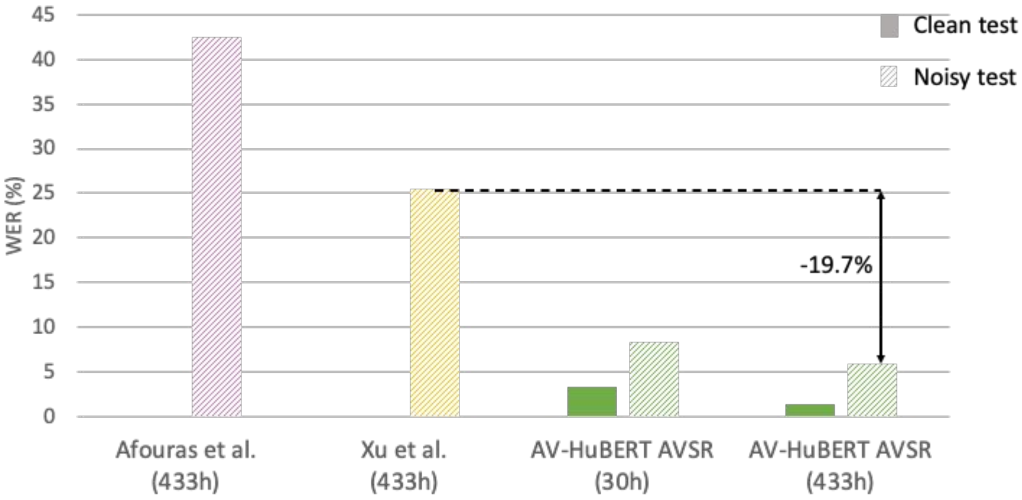
AV-HuBERT results of AVSR



- 35.1% absolute WER reduction
- beats 433hr-supervised model on with just 30hr labeled
- More robust to noise
  - 3.3% on clean, 7.8% on noisy

- Improve with more labeled data
  - 1.4% on clean, 5.8% on noisy

Comparison with Prior Work



- 19.7% absolute WER reduction compared to the prior SOTA

Conclusion

- Self-supervised learning is also very effective for audio-visual speech
- Multimodal self-supervised learning can benefit unimodal downstream tasks
- Visual information and self-supervised learning makes speech recognition more robust

Code and models available at https://facebookresearch.github.io/av_hubert

Thank you