JUNE 12, 2022

VISUALLY-ENHANCED SELF-SUPERVISED SPEECH MODELS

DAVID HARWATH
Assistant Professor, UTCS

The University of Texas at Austin
Department of Computer Science
College of Natural Sciences

Team

PI: David Harwath (UT), Hung-yi Lee (NTU)

Students at JHU: Layne Berry (UT), Elizabeth Boroda (JHU)

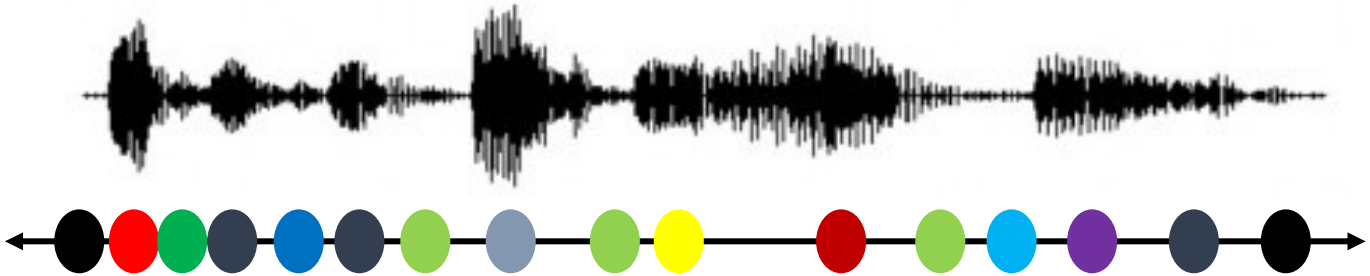
Remote students: Ian Shih (NTU), Hsuan-Fu Wang (NTU), Heng-Jui Chang (NTU), Puyuan Peng (UT)

Outline

1. Introduction to Visually-Grounded Speech Processing
2. Overview of what we're working on for the workshop
3. Students (Ian, Jeff, Layne) present our ongoing work

The Automatic Speech Recognition Learning Paradigm

- The training paradigm for speech recognition is >40 years old
  - {Speech, words} pairs enable alignment at phone/character level
  - Training becomes an exercise in aligning “beads on a string”



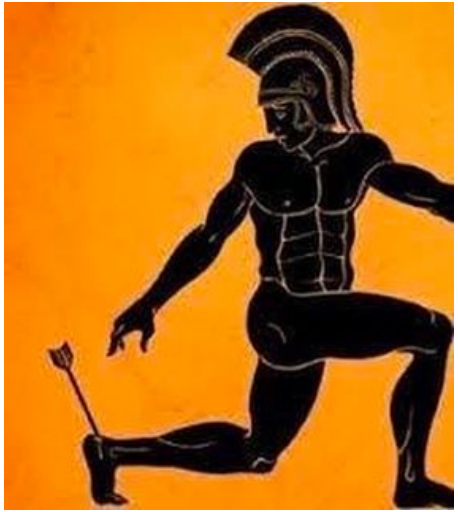
  - *This is not how humans learn speech!*

- Cost of annotations limits ASR to major languages of the world

- An ability to learn 1) with weakly constrained inputs from 2) freely available data, will be a major paradigm shift for ASR

A Fundamental Challenge for AI

- Many successful machine learning tasks rely on large quantities of ***annotated*** training data
  - Annotated data comes in {Input, Output} pairs

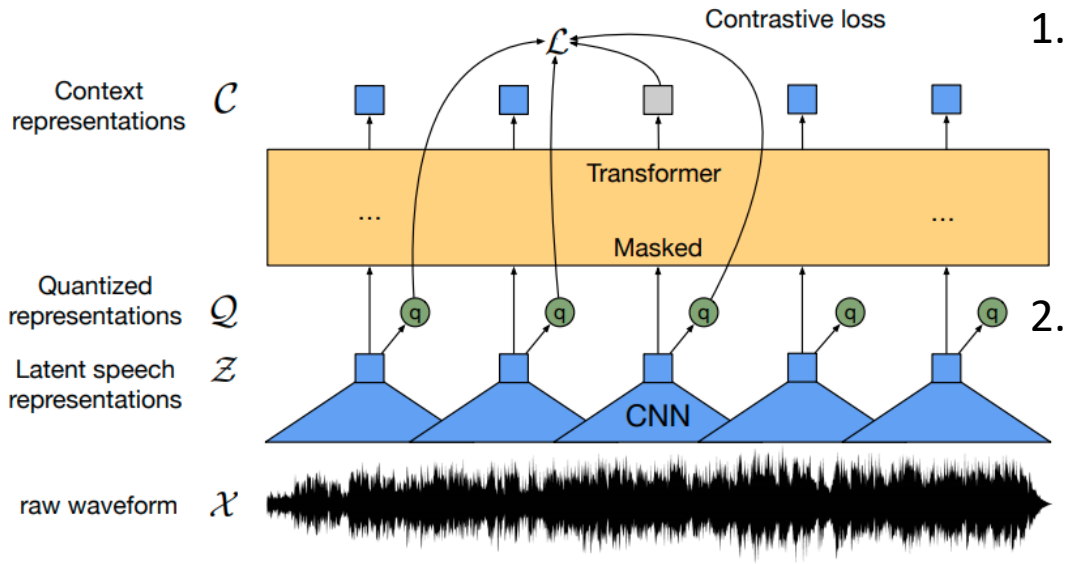


- Issues:
  - Training data should match the “testing” conditions
  - Annotating large corpora is time-consuming and expensive

- Challenge:
  - There is far more raw data in the world than annotated data
  - Can we build models that learn with much less supervision?

Self-Supervision to the Rescue?

Baevski et al., “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” 2020



1. Pre-train on **large** amount of **untranscribed** speech data (e.g. 1 to 60k hours) with masked language modeling objective

2. Add a projection layer on output + do supervised fine-tuning (e.g. with CTC) on **smaller** amount of **transcribed** speech

Hsu et al., “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,” 2021

Baevski et al., “data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language,” 2022

Are there other forms of SSL?

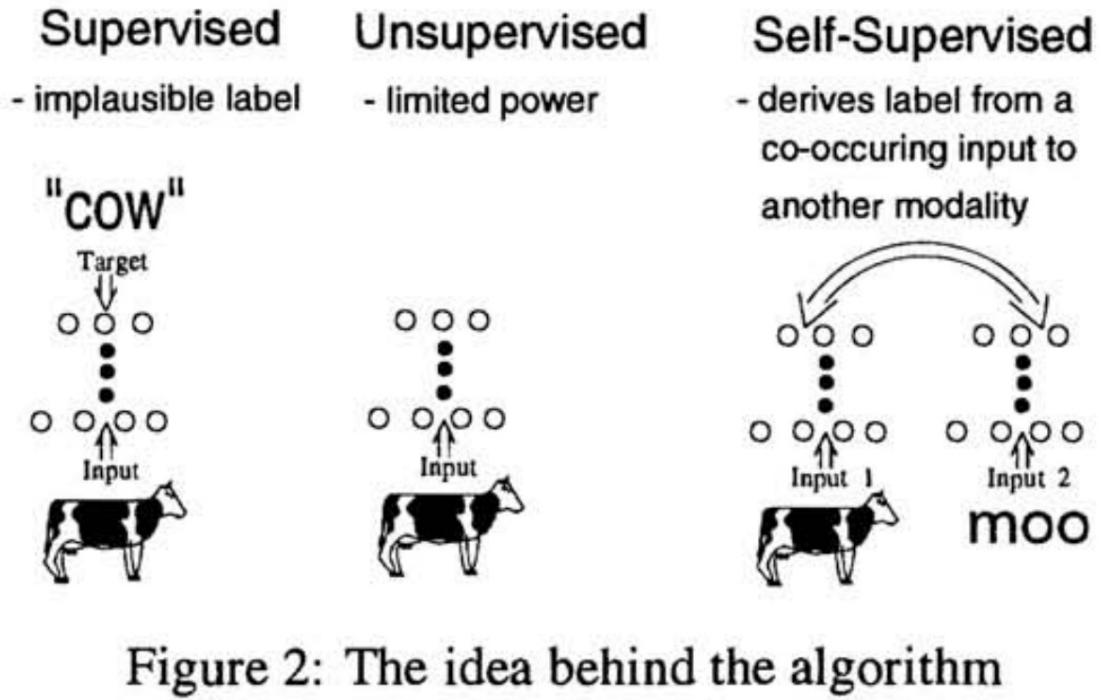
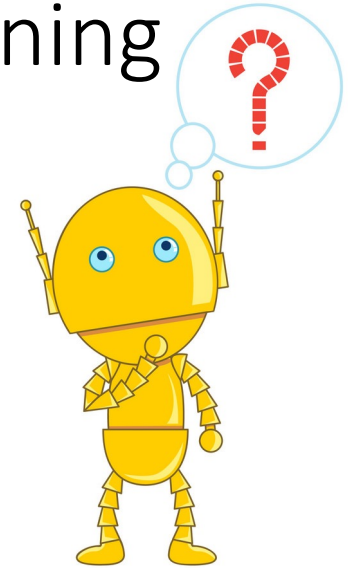


Figure 2: The idea behind the algorithm

Virginia de Sa, "Learning Classification with Unlabeled Data," Proc. NeurIPS 1994

Grounding as a learning objective

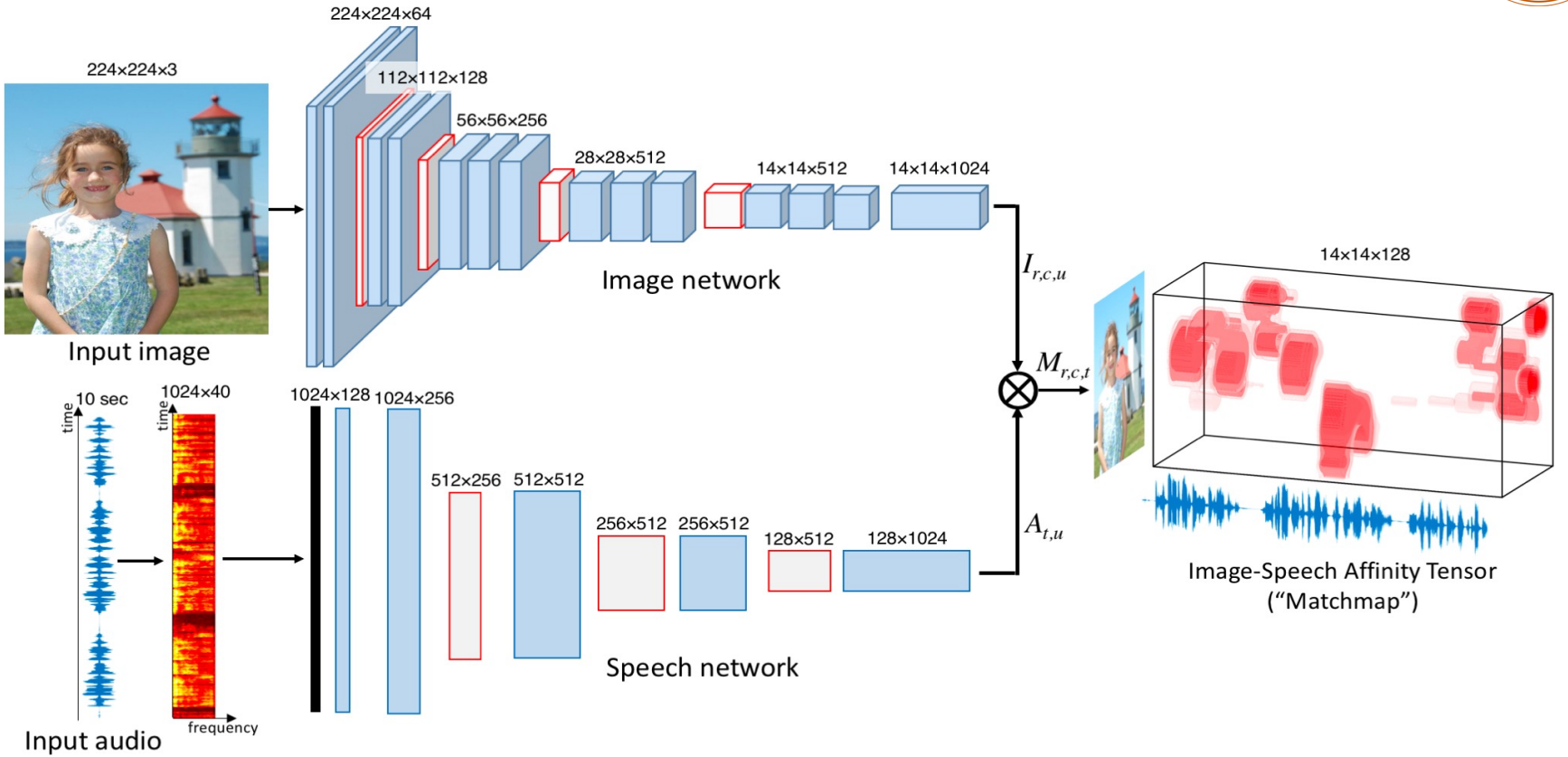


If a model can associate the speech waveform that it hears with the visual images that it sees, then it must have implicitly learned to:

1) Recognize words in speech
2) Recognize objects in images
3) Ground words to objects

Data: Spoken audio captions for MIT Places

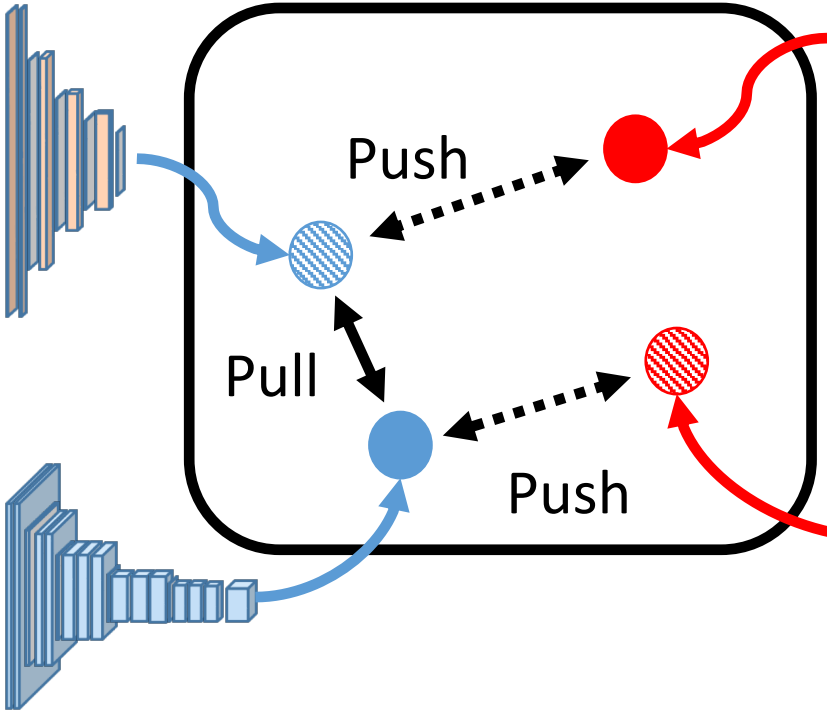
Jointly Embedding Speech and Images

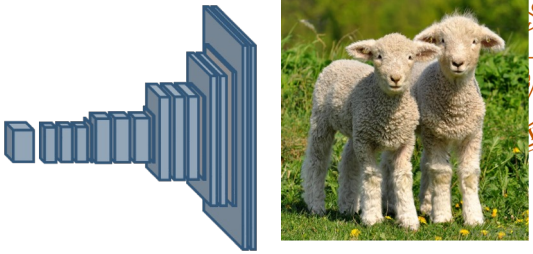


Training with a contrastive loss

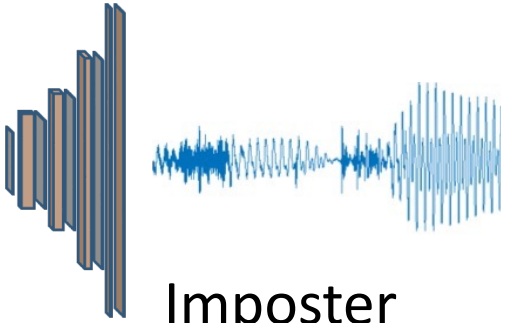
Paired example p

Anchor example a





Imposter example j

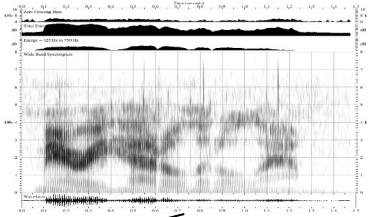


Imposter example i

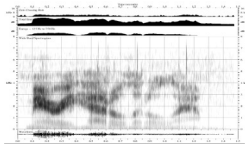
Various formulations of contrastive loss have been used, e.g. triplet [Harwath et al., 2016] and InfoNCE [Ilharco et al., 2019]

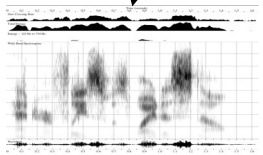
Evaluation: image and caption retrieval

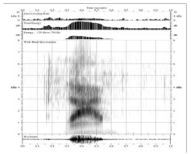
Image Retrieval:
Given caption, find image



Caption Retrieval:
Given image, find caption

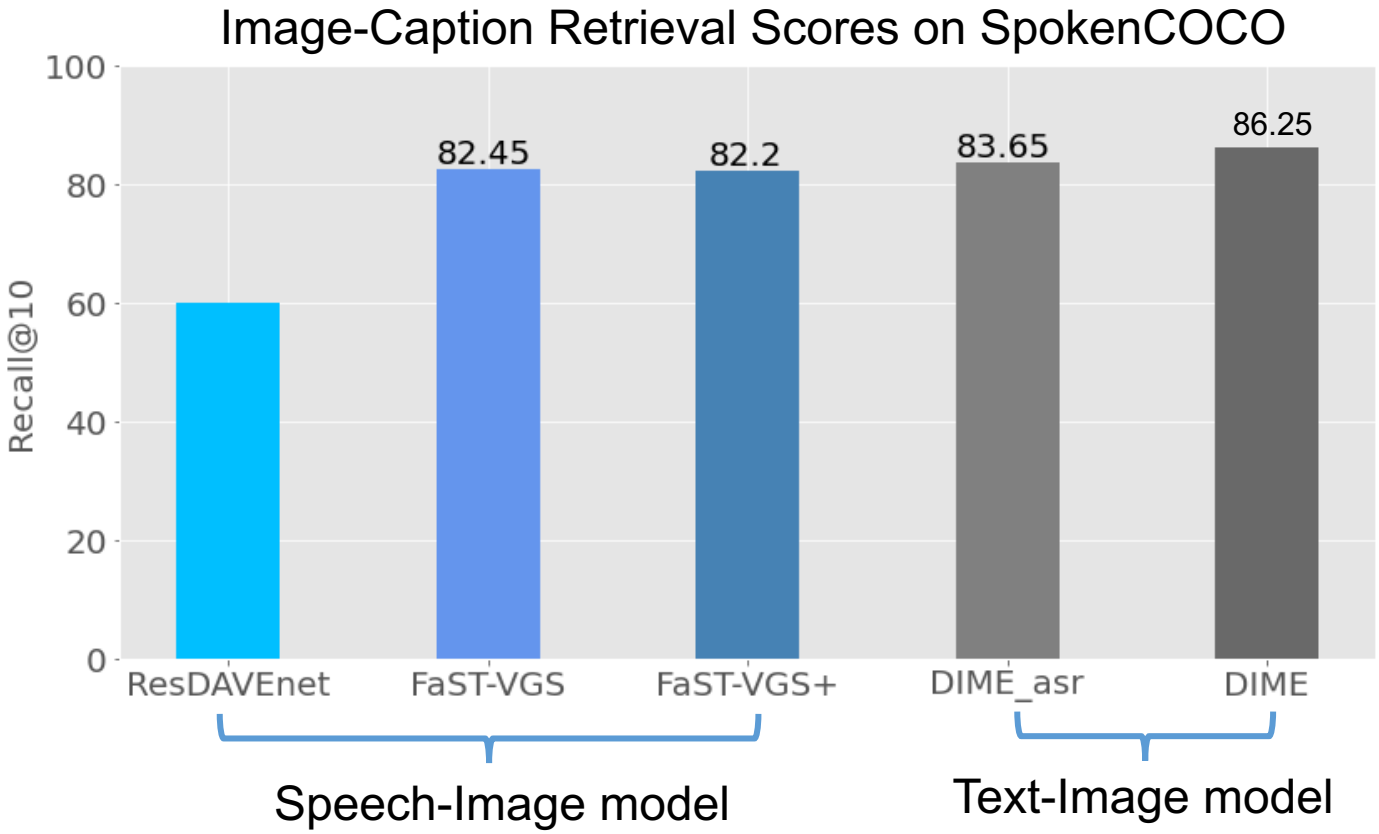






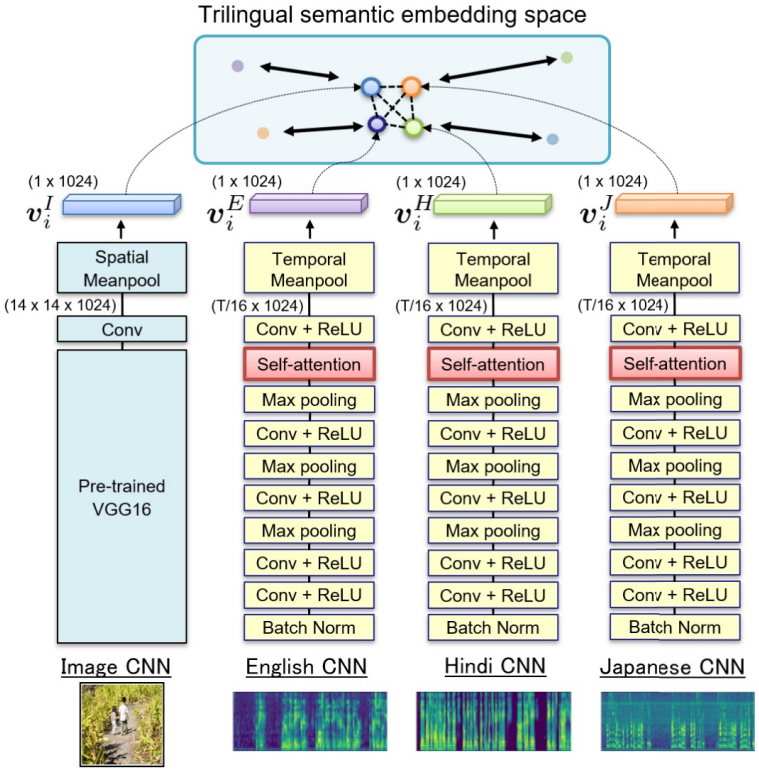
Evaluation metric: P(correct result is in top 10 retrieved examples)
(Recall @ 10)

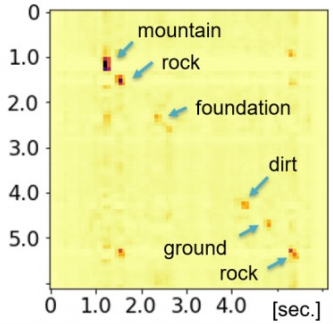
Speech-Image Retrieval

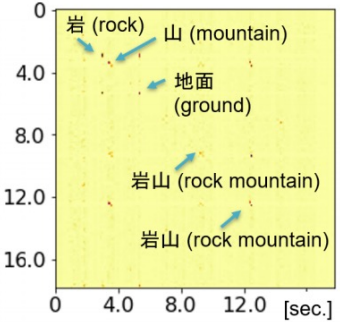



flybe.
A1 23

Vision as a “Rosetta Stone”







English caption: “*A **mountain rock foundation** with a lot of **dirt** on the **ground** and **rock***”

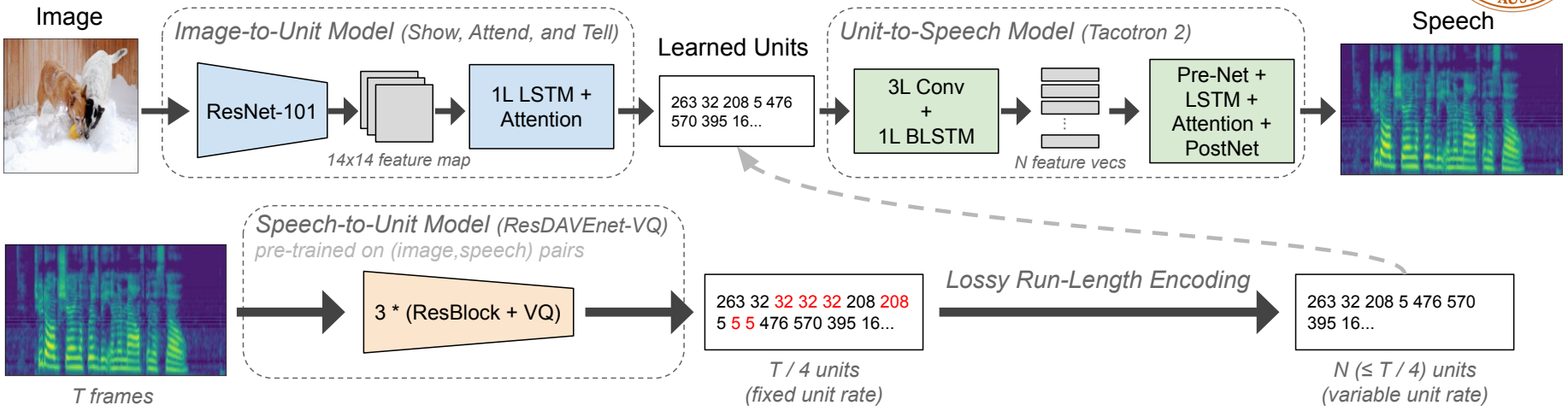
Japanese caption: “*真っ青な空の下に白い**岩山**がある。**地面**も真っ白で、波のような形をした大きな**岩山**が連なっていて、さらに大きな**岩山**が上のほうに乗り出していて複雑な形になっている*”

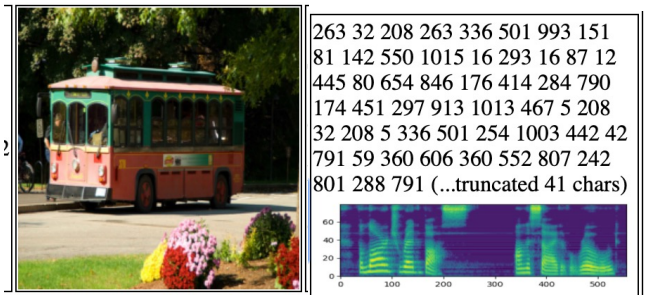
horses (0.75); घोड़े:the horses (0.66)

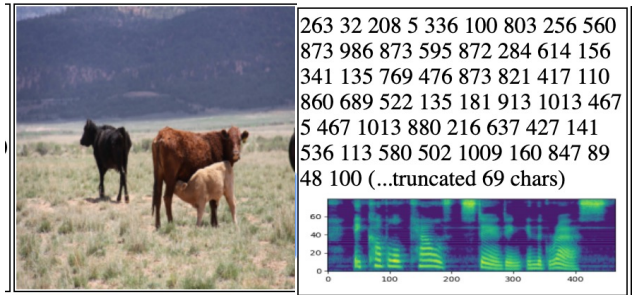
trees (0.83); पेड़:the trees (0.75)

Ohishi, Kimura, Kawanishi, Kashino, Harwath, and Glass, “Trilingual Semantic Embeddings of Visually Grounded Speech with Self-Attention Mechanisms,” ICASSP 2020
Azuh, Harwath, and Glass, “Towards Bilingual Lexicon Discovery from Visually Grounded Speech Audio,” Interspeech 2019

Generating spoken captions *without text!*

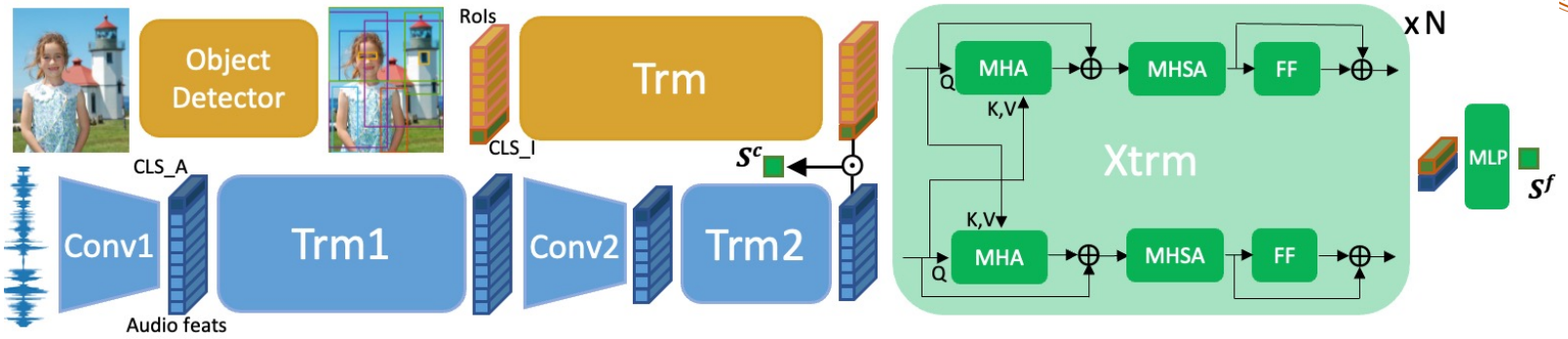


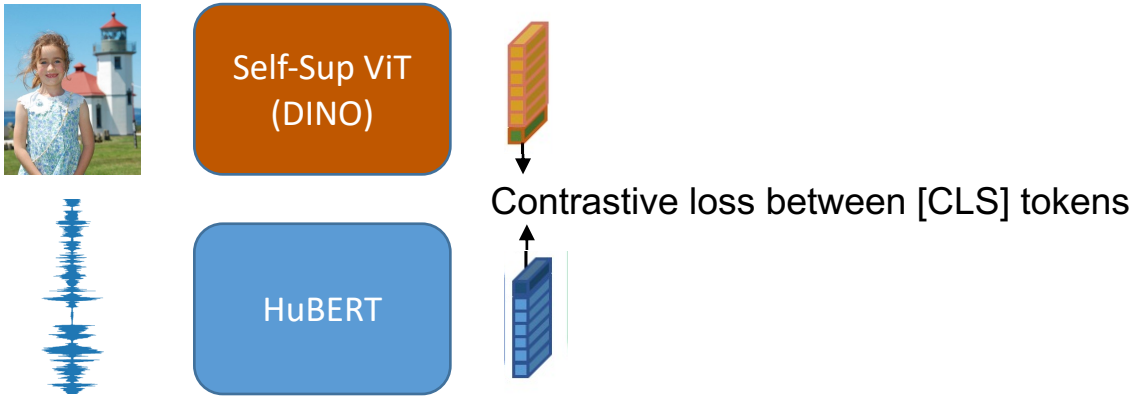




Harwath, Hsu, and Glass, "Learning Hierarchical Discrete Linguistic Units from Visually Grounded Speech," ICLR 2020
Hsu, Harwath, and Glass, "Text-Free Image-to-Speech Synthesis Using Learned Segmental Units," NeurIPS SAS Workshop 2020

FaST-VGS and VG-HuBERT

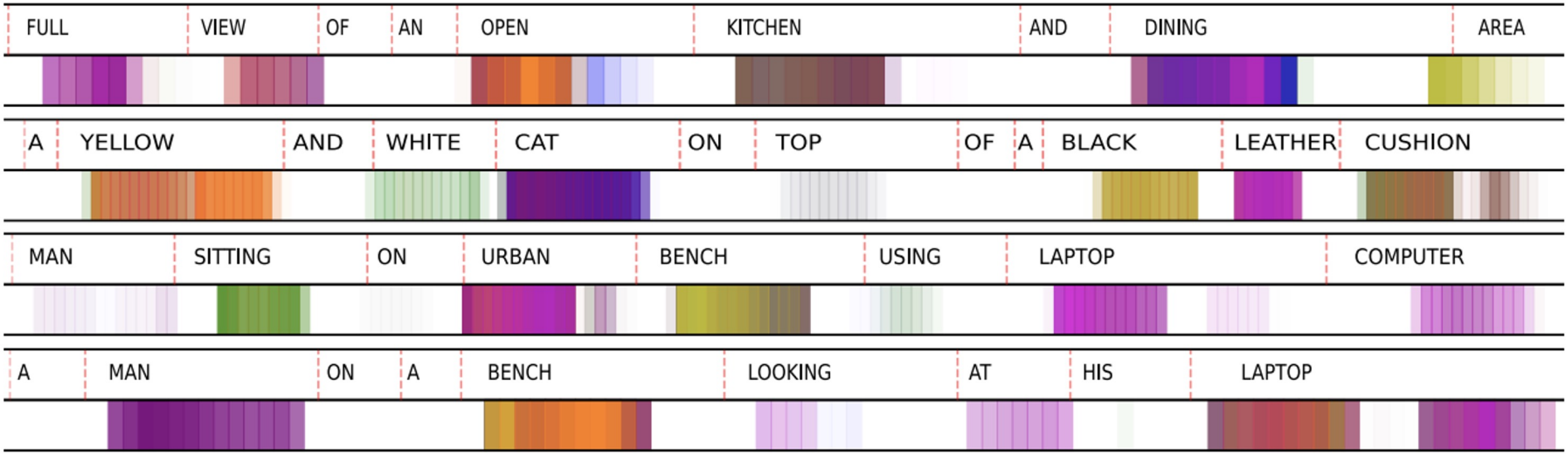




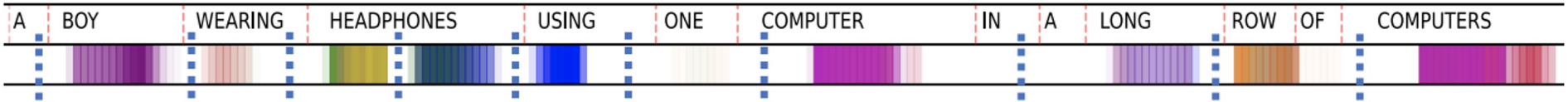
Results on SUPERB

Method	#Params	Data	Speaker			Content					Semantics			ParaL
			SID	ASV	SD	PR	ASR (WER)		KS	QbE	IC	SF		ER
			Acc ↑	EER ↓	DER ↓	PER ↓	w/o ↓	w/LM ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	15.21	8.63	0.0058	9.10	69.64	52.94	35.39
PASE+	7.83M	LS50	37.99	11.61	8.68	58.87	25.11	16.62	82.54	0.0072	29.82	62.14	60.17	57.86
APC	4.11M	LS360	60.42	8.56	10.53	41.98	21.28	14.74	91.01	0.0310	74.69	70.46	50.89	59.33
VQ-APC	4.63M	LS360	60.15	8.72	10.45	41.08	21.20	15.21	91.11	0.0251	74.48	68.53	52.91	59.66
NPC	19.38M	LS360	55.92	9.40	9.34	43.81	20.20	13.91	88.96	0.0246	69.44	72.79	48.44	59.08
Mockingjay	85.12M	LS360	32.29	11.66	10.54	70.19	22.82	15.48	83.67	6.6E-04	34.33	61.59	58.89	50.28
TERA	21.33M	LS360	57.57	15.89	9.96	49.17	18.17	12.16	89.48	0.0013	58.42	67.50	54.17	56.27
wav2vec	32.54M	LS960	56.56	7.99	9.9	31.58	15.86	11.00	95.59	0.0485	84.92	76.37	43.71	59.79
vq-wav2vec	34.15M	LS960	38.80	10.38	9.93	33.48	17.71	12.80	93.38	0.0410	85.68	77.68	41.54	58.24
wav2vec 2.0 Base	95.04M	LS960	75.18	6.02	6.08	5.74	6.43	4.79	96.23	0.0233	92.35	88.30	24.77	63.43
HuBERT Base	94.68M	LS960	81.42	5.11	5.88	5.41	6.42	4.97	96.30	0.0736	98.34	88.53	25.20	64.92
FaST-VGS	187.87M	LS960+SC742	41.49	6.54	6.50	16.30	13.46	9.51	96.85	0.0546	98.37	84.91	32.33	57.37
FaST-VGS+	217.23M	LS960+SC742	41.34	5.87	6.05	7.76	8.83	6.37	97.27	0.0562	98.97	88.15	27.12	60.96
modified CPC	1.84M	LL60k	39.63	12.86	10.38	42.54	20.18	13.53	91.88	0.0326	64.09	71.19	49.91	60.96
WavLM Base+	94.70M	Mix94k	86.84	4.26	4.07	4.07	5.64	—	96.69	**0.0990**	**99.16**	89.73	21.54	67.98
wav2vec 2.0 Large	317.38M	LL60k	86.14	5.65	5.62	4.75	3.75	3.10	96.66	0.0489	95.28	87.11	27.31	65.64
HuBERT Large	316.61M	LL60k	90.33	5.98	5.75	3.53	3.62	2.94	95.29	0.0353	98.76	89.81	21.76	67.62
WavLM Large	316.62M	Mix94k	**95.25**	**4.04**	**3.47**	**3.09**	**3.51**	—	**97.40**	0.0827	99.10	**92.25**	**17.61**	**70.03**

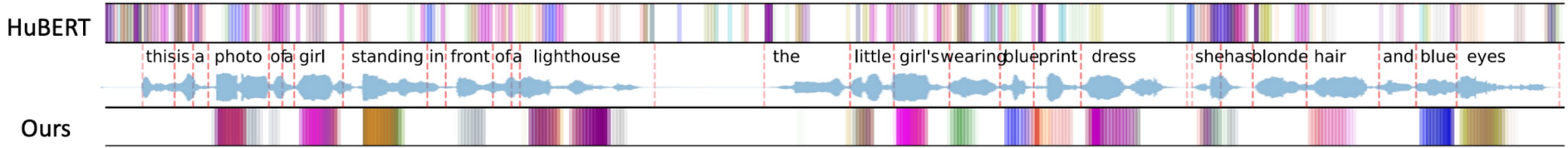
Detecting and Segmenting Words

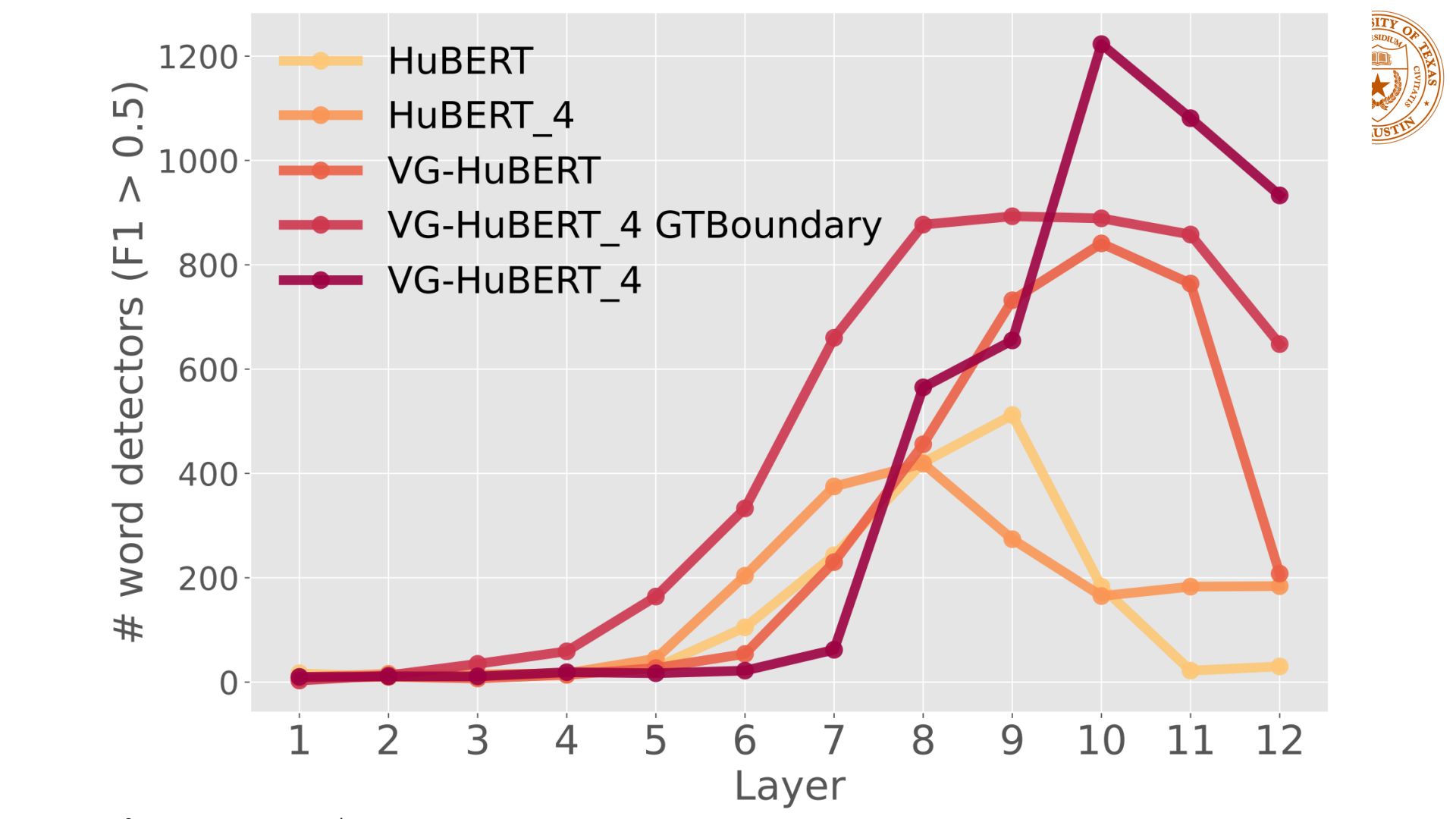


Generate word boundaries: midpoints of adjacent boundaries of attention segments:



This same word segmentation ability isn’t present in HuBERT




HuBERT
HuBERT_4
VG-HuBERT
VG-HuBERT_4 GTBoundary
VG-HuBERT_4
word detectors (F1 > 0.5)
0
200
400
600
800
1000
1200
1
2
3
4
5
6
7
8
9
10
11
12
Layer

Word Discovery on SpokenCOCO

	Area				Boundary					Word	
Model	WC	tIoU	CD (ms)	A-score	Precision	Recall	F_1	OS	R-val	Purity	WD
ResDAVEnet-VQ [24]	98.39	23.78	119.4	38.30	10.42	50.96	17.30	3.8883	-2.5077	31.6 ± 0	262 ± 0
W2V2 [9]	89.64	31.06	106.7	46.13	11.88	24.79	16.06	1.0868	-0.3110	42.9 ± 0.2	396 ± 15
HuBERT [10]	89.90	31.62	105.8	46.78	11.90	25.81	16.29	1.1682	-0.3672	45.3 ± 0.3	500 ± 9
FaST-VGS [12]	75.92	54.86	67.8	63.69	28.99	26.17	<u>27.51</u>	-0.0972	<u>0.4010</u>	70.9 ± 0.5	1011 ± 12
FaST-VGS+ [27]	87.36	47.49	72.2	61.53	22.66	27.86	24.99	0.2293	0.2854	72.3 ± 0.3	1026 ± 12
VG-W2V2	74.24	41.09	79.1	52.90	18.47	19.78	19.10	0.0709	0.2886	62.5 ± 0.4	743 ± 14
VG-W2V2$_4$	71.28	54.06	82.0	61.49	28.15	22.90	25.26	-0.1864	0.3967	73.0 ± 0.3	<u>1182</u> ± 20
VG-W2V2$_5$	73.32	53.09	72.2	61.58	28.70	25.45	26.98	-0.1132	0.3994	75.4 ± 0.1	1149 ± 16
VG-HuBERT	73.16	44.02	89.0	54.97	18.31	18.90	18.60	0.0326	0.2960	63.2 ± 0.4	823 ± 20
VG-HuBERT$_3$	74.91	57.23	60.6	**64.89**	35.90	27.03	**30.84**	-0.2472	**0.4442**	75.3 ± 0.2	1167 ± 26
VG-HuBERT$_4$	73.97	56.35	78.9	<u>63.97</u>	28.39	25.64	26.94	-0.0970	0.3964	75.2 ± 0.2	**1230** ± 18

ZeroSpeech Spoken Term Discovery

	Words	NED	Cov	M-score	Prec.	Rec.	F_1
J.V. [28]	18821	**32.4**	7.9	14.1	32.1	3.2	5.9
G.S.[8]	92544	72.3	76.8	40.7	27.8	45.5	34.5
R.S. (2017) [29]	**321603**	52.5	28.7	35.8	25.8	29.9	27.7
ES-KMeans [4]	42473	72.3	**100.0**	43.4	39.6	61.4	48.2
SEA [30]	240033	89.5	99.5	19.0	27.3	75.9	40.1
PDTW [31]	85425	48.2	85.4	64.5	26.5	**88.2**	40.8
VG-HuBERT$_3$ (Ours)	104696	42.5	95.4	**71.8**	**44.7**	54.2	**49.0**

Buckeye Word Segmentation

Model	Boundary				Token
	Prec.	Rec.	F_1	R-val.	F_1
Adaptor gram. [32]	15.9	**57.7**	25.0	-139.9	4.4
SylSeg [33]	27.7	28.9	28.3	37.7	19.3
ES-KMeans [4]	30.3	16.6	21.4	39.1	19.2
BES-GMM [34]	31.5	12.4	17.8	37.2	18.6
SCPC [35]	36.9	29.9	33.0	45.6	-
mACPC [36]	<u>42.1</u>	30.3	35.1	<u>47.4</u>	-
DPDP [37]	35.3	37.7	<u>36.4</u>	44.3	<u>25.0</u>
VG-HuBERT$_3$ (Ours)	**47.6**	<u>42.3</u>	**44.8**	**54.2**	**31.0**

Research directions for this workshop

- Previously explored: Parallel SpeechCLIP
  - Main research idea: can we use the CLIP image encoder as a teacher model for a speech encoder like HuBERT to improve its results on SUBERB tasks
  - Not very promising results (showed at last meeting)
- Currently exploring: Cascaded SpeechCLIP (Ian, Jeff, and Layne will present this today)
  - Main research idea: can we use the CLIP image *and text* encoders as a teacher model for a speech encoder like HuBERT?
  - Downstream tasks: SUPERB tasks, Unsupervised ASR, Unsupervised Speech Translation
- Using VG-HuBERT's word discovery ability for SSL
  - Use automatic word segmentation for non-uniform downsampling of speech signal
  - Use discovered word segments/clusters as masks/targets for HuBERT training
  - Use discovered words for word-level unsupervised ASR (no phonemization needed)

Pre-trained models we're using

- CLIP
- HuBERT
- FaST-VGS
    - Current SotA for speech-image retrieval
    - Code and checkpoints already available to team
- VG-HuBERT
    - Version of HuBERT model further trained to ground ~600k spoken caption waveforms to the images they describe
    - Achieves SotA unsupervised word segmentation results
    - Code and checkpoints already made available to the team

Speech-Image Datasets Available

- Flickr8k
  - 40k spoken descriptions (read speech) of 8k Flickr8k images
- Places Audio (English)
  - 400k spoken descriptions (spontaneous speech) of 400k MIT Places images
- SpokenCOCO
  - 600k spoken descriptions (read speech) of the MSCOCO captions for 125k images
- Places Audio (Hindi)
  - 100k spoken descriptions (Hindi spontaneous speech) of 100k MIT Places images
- Spoken Moments
  - 500k spoken descriptions (spontaneous speech) of 3 second, action-oriented videos

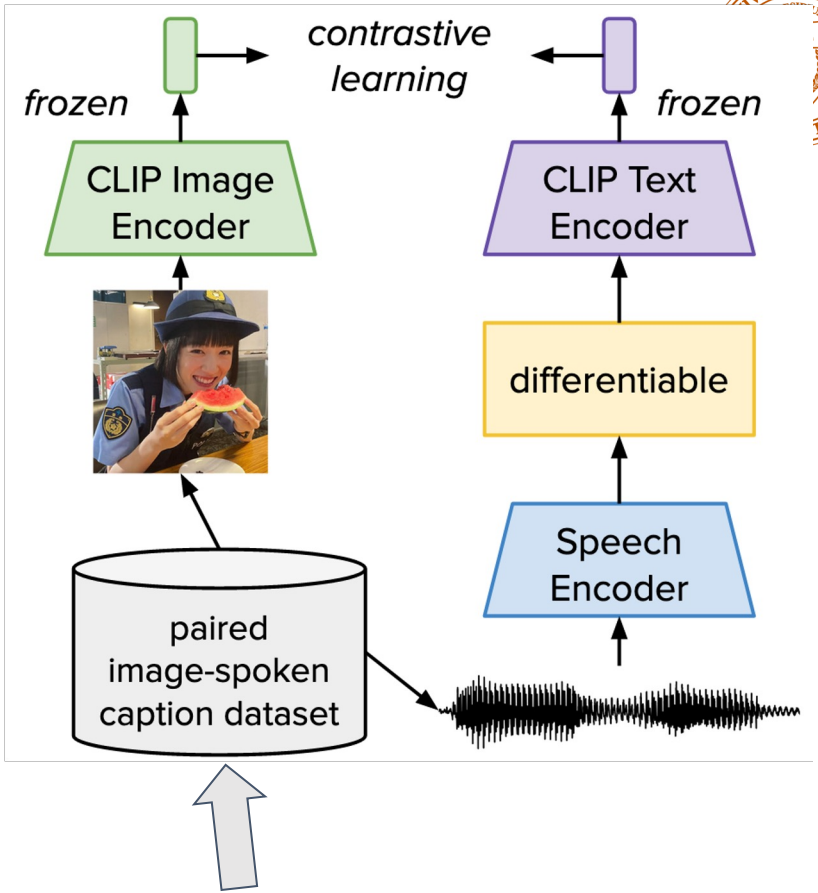
Cascaded SpeechCLIP

Main idea: can the CLIP model be used to help train a speech encoder?

We can use the learned representations from the speech encoder on downstream tasks like SUPERB, ZeroSpeech

This model also opens the door to some new possibilities:

- Unsupervised speech recognition: Learn to map English speech to English text using the image as a guide
- Unsupervised speech-to-text translation: Use Hindi input speech, learn to map to English text using the image as a guide



This can be English (SpokenCOCO, Places Audio) or Hindi (Places Audio Hindi)