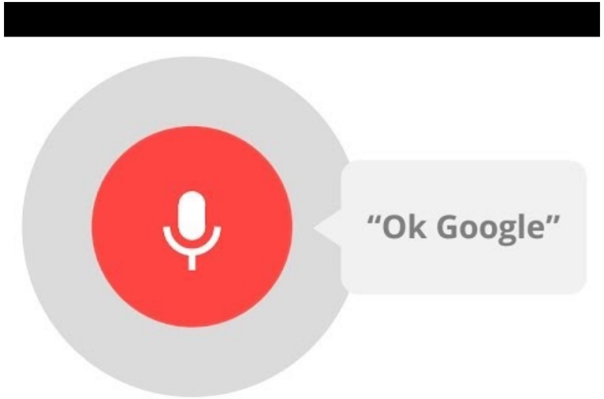
NOVEMBER 8, 2022

VISUALLY GROUNDING SPEECH FOR MULTIMEDIA RETRIEVAL AND BEYOND

DAVID HARWATH
Assistant Professor, UTCS

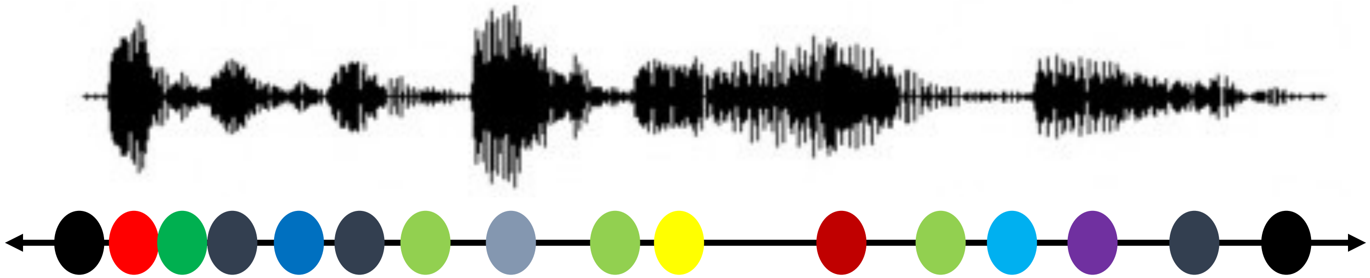
The University of Texas at Austin
Department of Computer Science
College of Natural Sciences

ASR: A ML Success Story



The Automatic Speech Recognition Learning Paradigm

- The traditional 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"



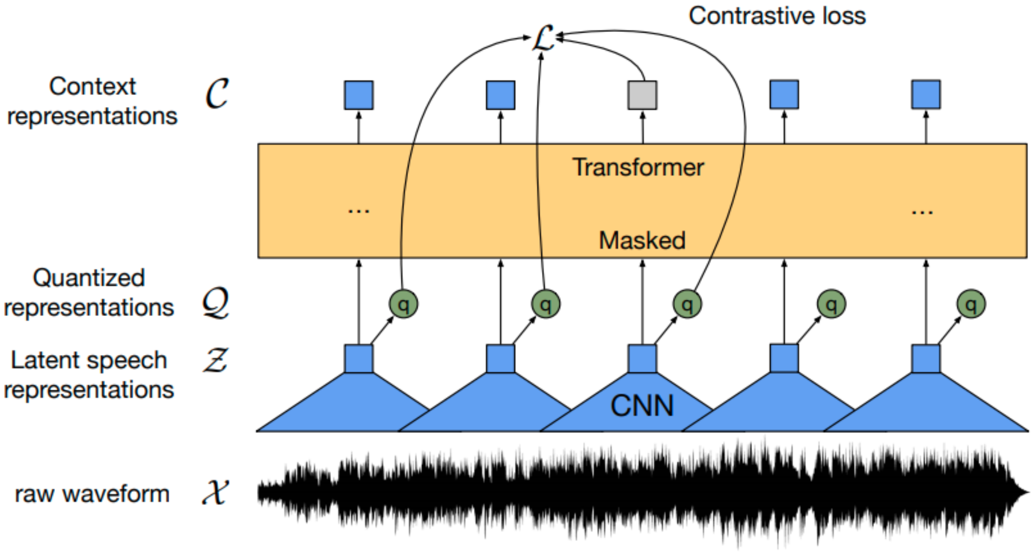
  - *<u>This is not how humans learn speech!</u>*

- 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

- 7,151 languages spoken worldwide today, half of which have less than 10,000 speakers each
- Approximately 3,000 languages are *unwritten*

Self-Supervised Learning (SSL) to the Rescue



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

Baevski et al., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," 2020
Hsu et al., "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," 2021

Wav2vec2.0 ASR on Librispeech

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

Table 2: WER on Librispeech when using all 960 hours of labeled data (cf. Table 1).

Model	Unlabeled data	LM	dev clean	dev other	test clean	test other
Supervised						
CTC Transf [51]	-	CLM+Transf.	2.20	4.94	2.47	5.45
S2S Transf. [51]	-	CLM+Transf.	2.10	4.79	2.33	5.17
Transf. Transducer [60]	-	Transf.	-	-	2.0	4.6
ContextNet [17]	-	LSTM	1.9	3.9	1.9	4.1
Conformer [15]	-	LSTM	2.1	4.3	1.9	3.9
1h labeled						
Discrete BERT [4]	LS-960	4-gram	8.5	16.4	9.0	17.6
BASE	LS-960	4-gram	5.0	10.8	5.5	11.3
		Transf.	3.8	9.0	4.0	9.3
LARGE	LS-960	Transf.	3.8	7.1	3.9	7.6
	LV-60k	Transf.	2.9	5.4	2.9	5.8

Contemporary models, fully supervised (**960 hours** of **transcribed** speech)

Wav2vec2.0 fine-tuned on **1 hour** of **transcribed** speech

Are there other forms of SSL?

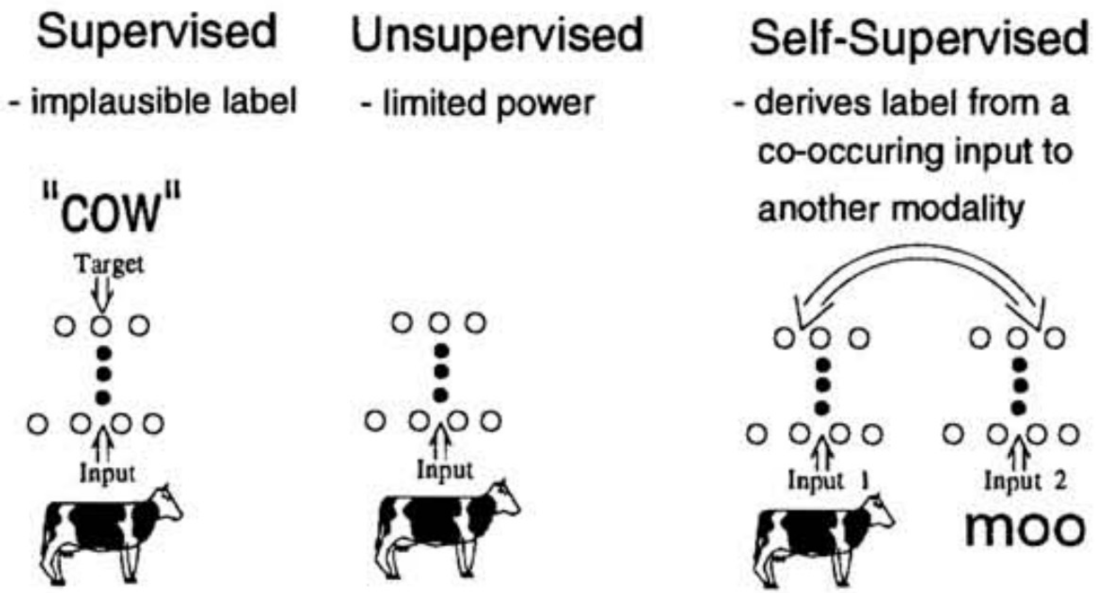
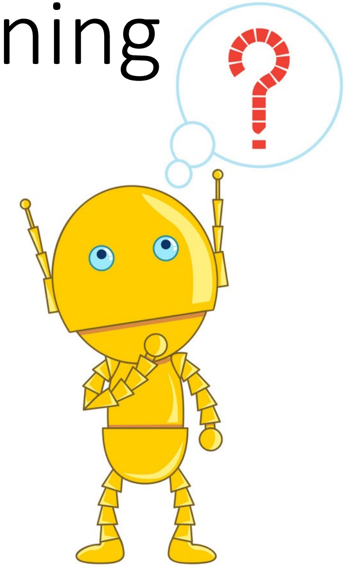


Figure 2: The idea behind the algorithm

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

Grounding as a learning objective



Learning to associate the speech you hear with the things you see…

…entails the ability to extract meaning from speech
…which entails the ability to recognize spoken word forms
…which entails the ability to recognize sub-word sounds

Talk Outline

1. Learning representations of speech with visual grounding [Harwath, Torralba, and Glass, NeurIPS 2016], [Harwath and Glass, ACL 2017], [Harwath et al., ECCV 2018]

2. Hybridizing dual-encoders and cross-modal attention models for visually grounding speech [Peng and Harwath, ICASSP 2022]

3. Emergent Word Discovery with Visually-Grounded HuBERT [Peng and Harwath, Interspeech 2022]

4. Learning audio-visual representations of instructional videos in the wild [Rouditchenko et al., Interspeech 2021]

5. Learning to generate spoken image descriptions without text [Hsu, Harwath, Miller, Song, and Glass, ACL 2021]

Talk Outline

1. Learning representations of speech with visual grounding [Harwath, Torralba, and Glass, NeurIPS 2016], [Harwath and Glass, ACL 2017], [Harwath et al., ECCV 2018]

2. Hybridizing dual-encoders and cross-modal attention models for visually grounding speech [Peng and Harwath, ICASSP 2022]

3. Emergent Word Discovery with Visually-Grounded HuBERT [Peng and Harwath, Interspeech 2022]

4. Learning audio-visual representations of instructional videos in the wild [Rouditchenko et al., Interspeech 2021]

5. Learning to generate spoken image descriptions without text [Hsu, Harwath, Miller, Song, and Glass, ACL 2021]

Visually Grounded Speech via Spoken audio captions

Instructions

This HIT is part of a MIT scientific research project. Your decision to complete this HIT is voluntary, and your responses are anonymous. The results of the research may be presented at scientific meetings, published in scientific journals, or made publicly available to other researchers. Clicking on the 'SUBMIT' button on the bottom of this page indicates that you are at least 18 years of age, you are a native English speaker, and you agree to complete this HIT voluntarily.

Connected

To complete this task, you must be:

- using a computer equipped with a microphone
- using the Chrome web browser
- in a relatively quiet environment

If your microphone is on and working, the volume meter at the right should move as you speak (after you grant permission for the site to use your microphone). Underneath the microphone volume meter you can see whether you are connected to server for recording. If you become disconnected, please continue recording after a connection is reestablished.

You will be presented with 4 image scenes. For each image, please:

- Press the [Record] button next to the image and then describe the image as if you were describing it to a blind person. During recording, the record button will be replaced with a stop button; end the recording by pressing the [Stop] button next to the image.
- After you record a caption, we will process the recording. If it is acceptable, it will be marked as [Great.] Otherwise, the sentence will be marked with a [Redo] and you must redo the recording of that sentence to complete the task.
- After all 3 descriptions have been accepted, the submit button at the bottom of the page will be enabled.

Here's an example of the level of detail we're looking for:

[Harwath, Torralba, and Glass, NeurIPS 2016]
[Harwath et al., ECCV 2018]

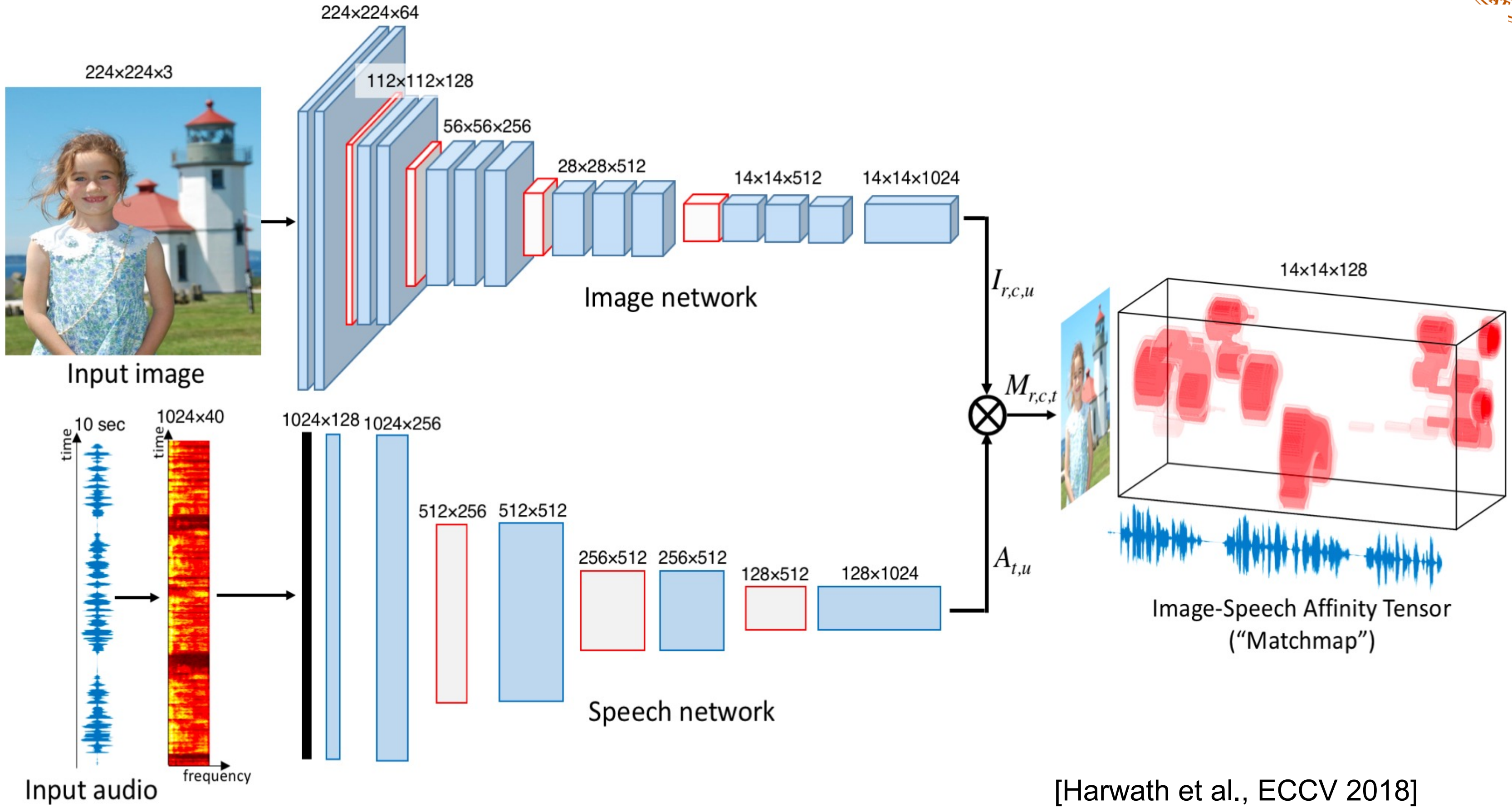
Spoken caption datasets we've collected

1. Flickr8k Audio Captions
   - 8,000 images from Flickr8k dataset [Hodosh et al., 2013] each with 5 text captions which are read aloud by native English speakers
2. Places Audio 400k
   - 400,000 images from MIT Places dataset [Zhou et al., 2014], each with one spoken English caption **(spontaneous speech)**
   - Approx. 100,000 of the images also have Hindi and Japanese captions
3. SpokenCOCO
   - 120,000 images from MSCOCO dataset [Lin et al., 2014], each with 5 text captions which are read aloud by native English speakers
4. Spoken Moments in Time
   - 500,000 short video clips from Moments in Time dataset [Monfort et al., 2019], each with one spoken English caption **(spontaneous speech)**

Spoken caption datasets we've collected

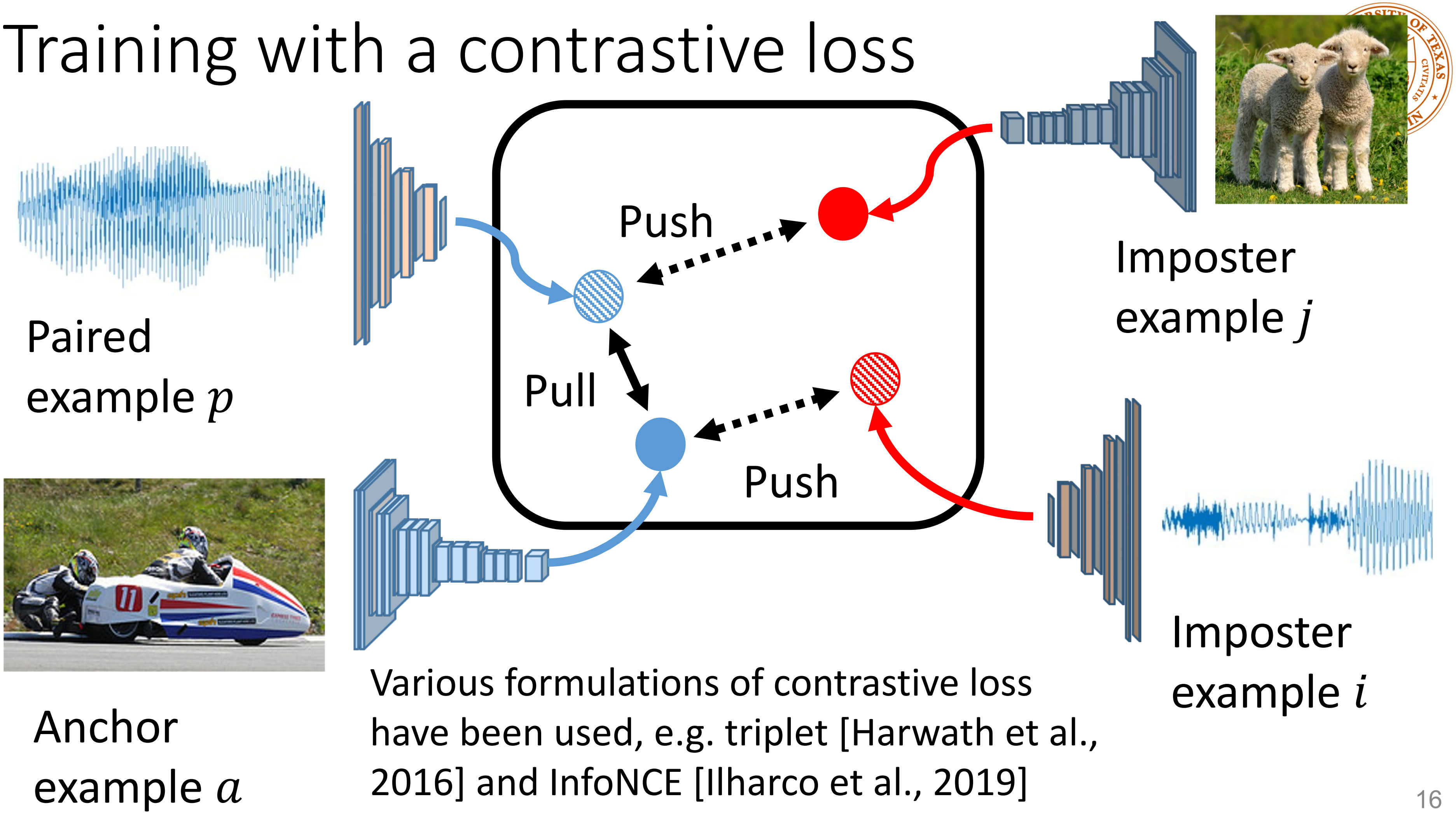
1. Flickr8k Audio Captions
   - 8,000 images from Flickr8k dataset [Hodosh et al., 2013] each with 5 text captions which are read aloud by native English speakers
2. Places Audio 400k
   - 400,000 images from MIT Places dataset [Zhou et al., 2014], each with one spoken English caption **(spontaneous speech)**
   - Approx. 100,000 of the images also have Hindi and Japanese captions
3. SpokenCOCO
   - 120,000 images from MSCOCO dataset [Lin et al., 2014], each with 5 text captions which are read aloud by native English speakers
4. Spoken Moments in Time
   - 500,000 short video clips from Moments in Time dataset [Monfort et al., 2019], each with one spoken English caption **(spontaneous speech)**

Jointly Embedding Speech and Images



[Harwath et al., ECCV 2018]

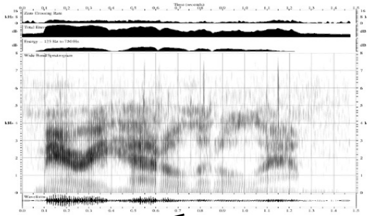
Training with a contrastive loss



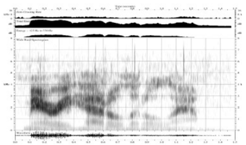
Various formulations of contrastive loss have been used, e.g. triplet [Harwath et al., 2016] and InfoNCE [Ilharco et al., 2019]

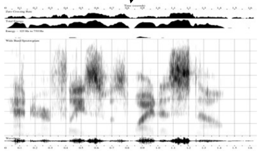
Evaluation: image and caption retrieval

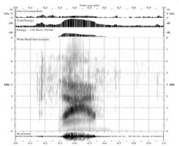
Image Retrieval:
Given caption, find image



Caption Retrieval:
Given image, find caption

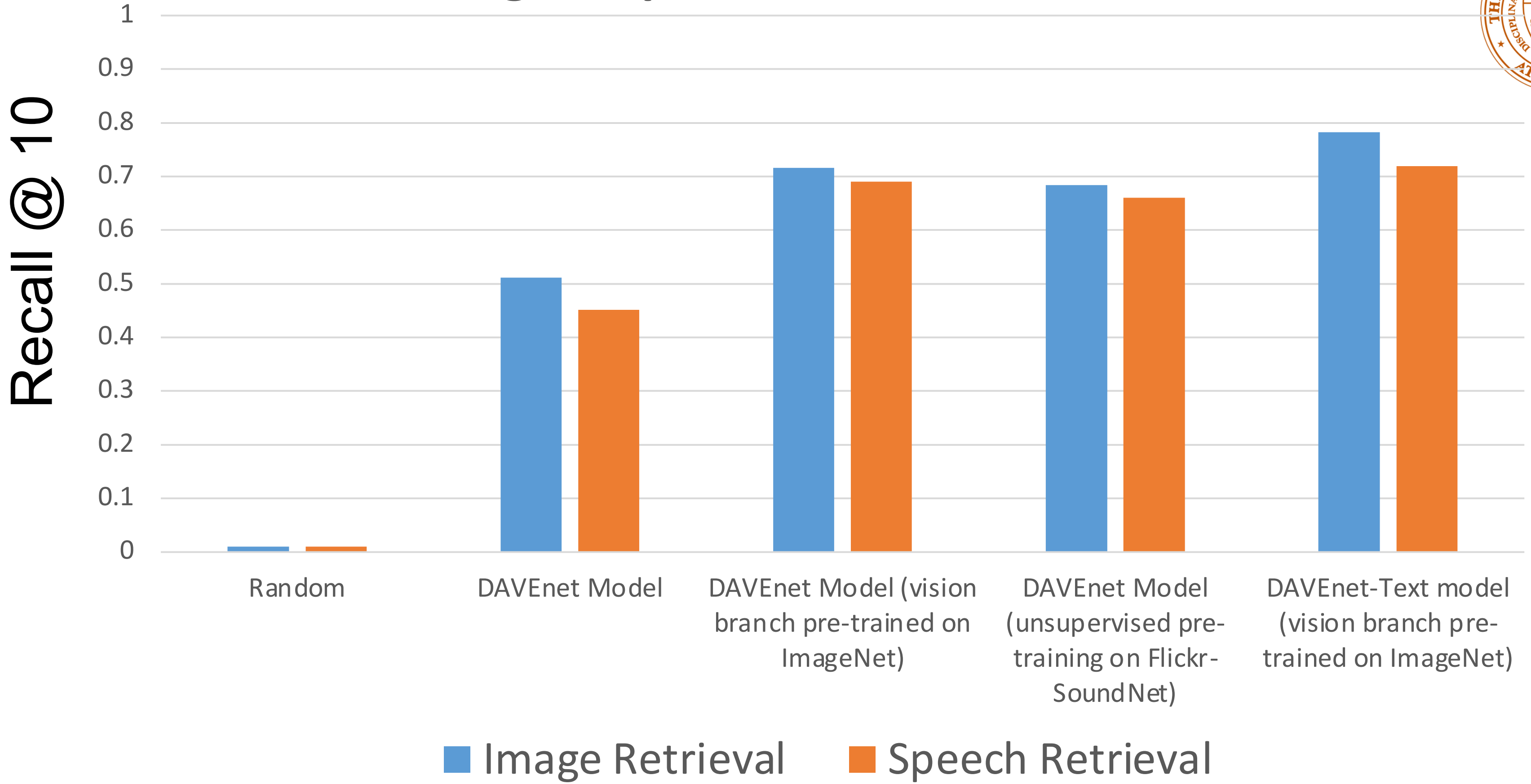






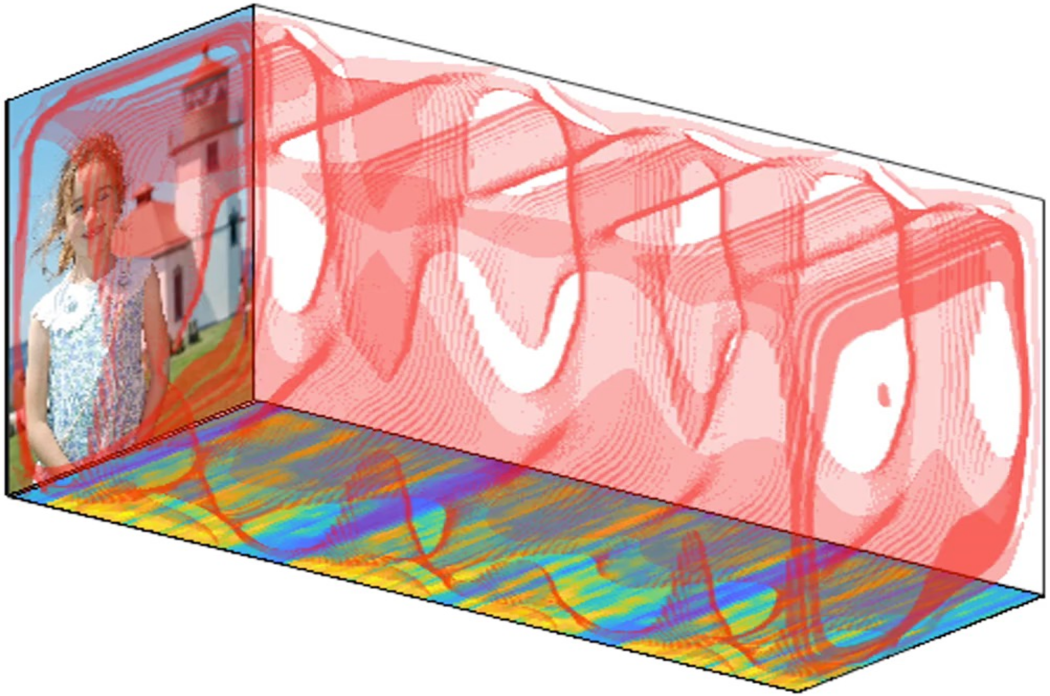
Evaluation metric: P(correct result is in top 10 retrieved examples)
(Recall @ 10)

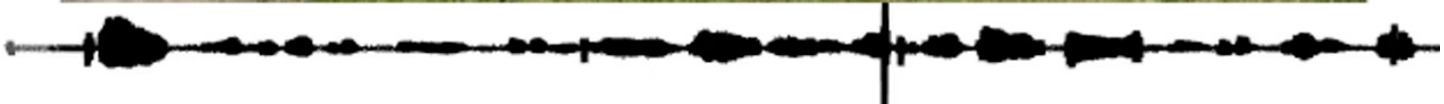
Image/Caption Retrieval on Places



Training set: 400k images + 400k captions; Testing set: 1k images + 1k captions

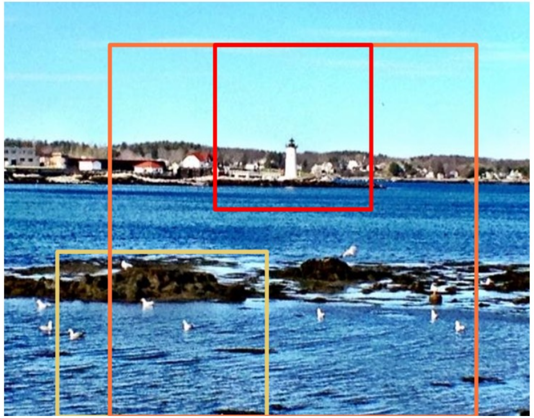
Matchmap convergence

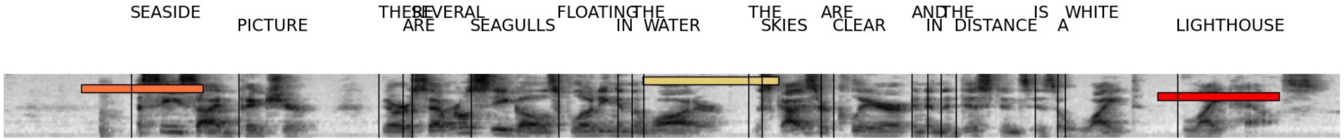


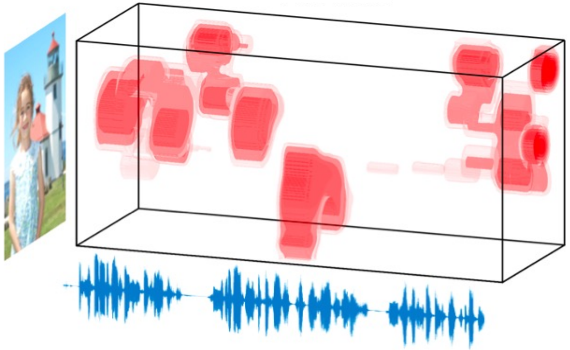


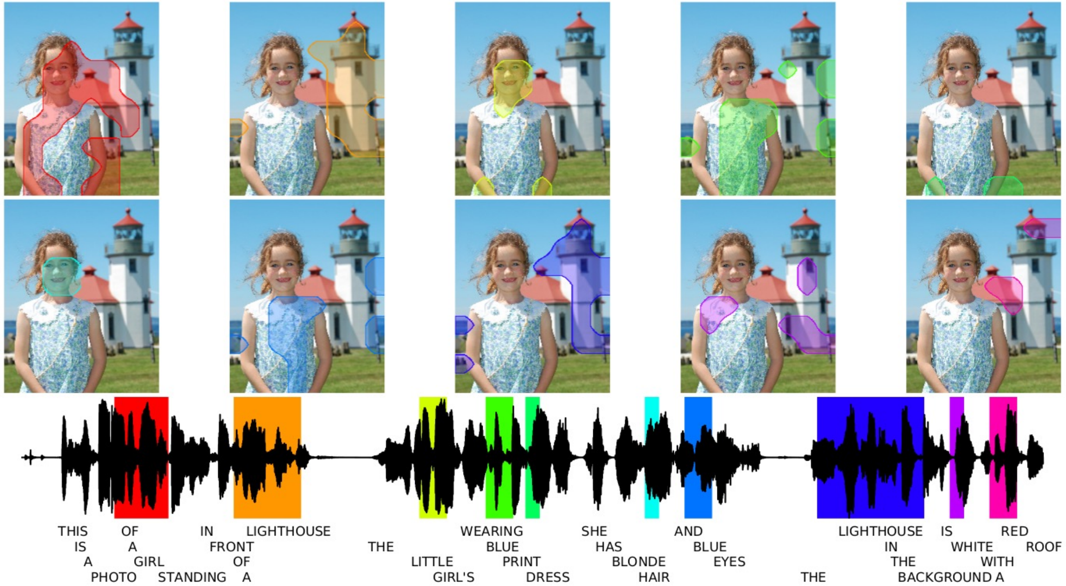

flybe.
A1
23

Semantic co-segmentation

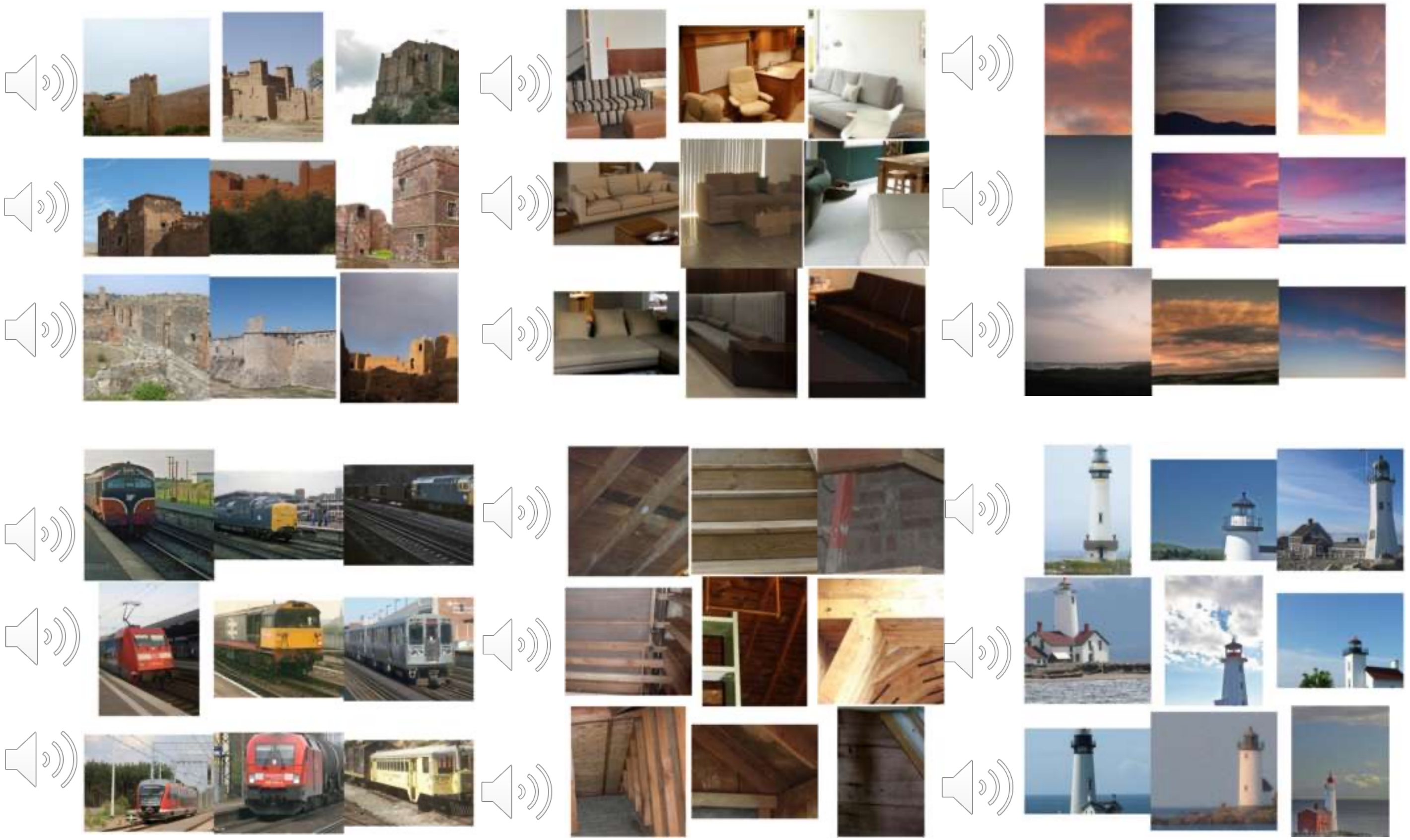


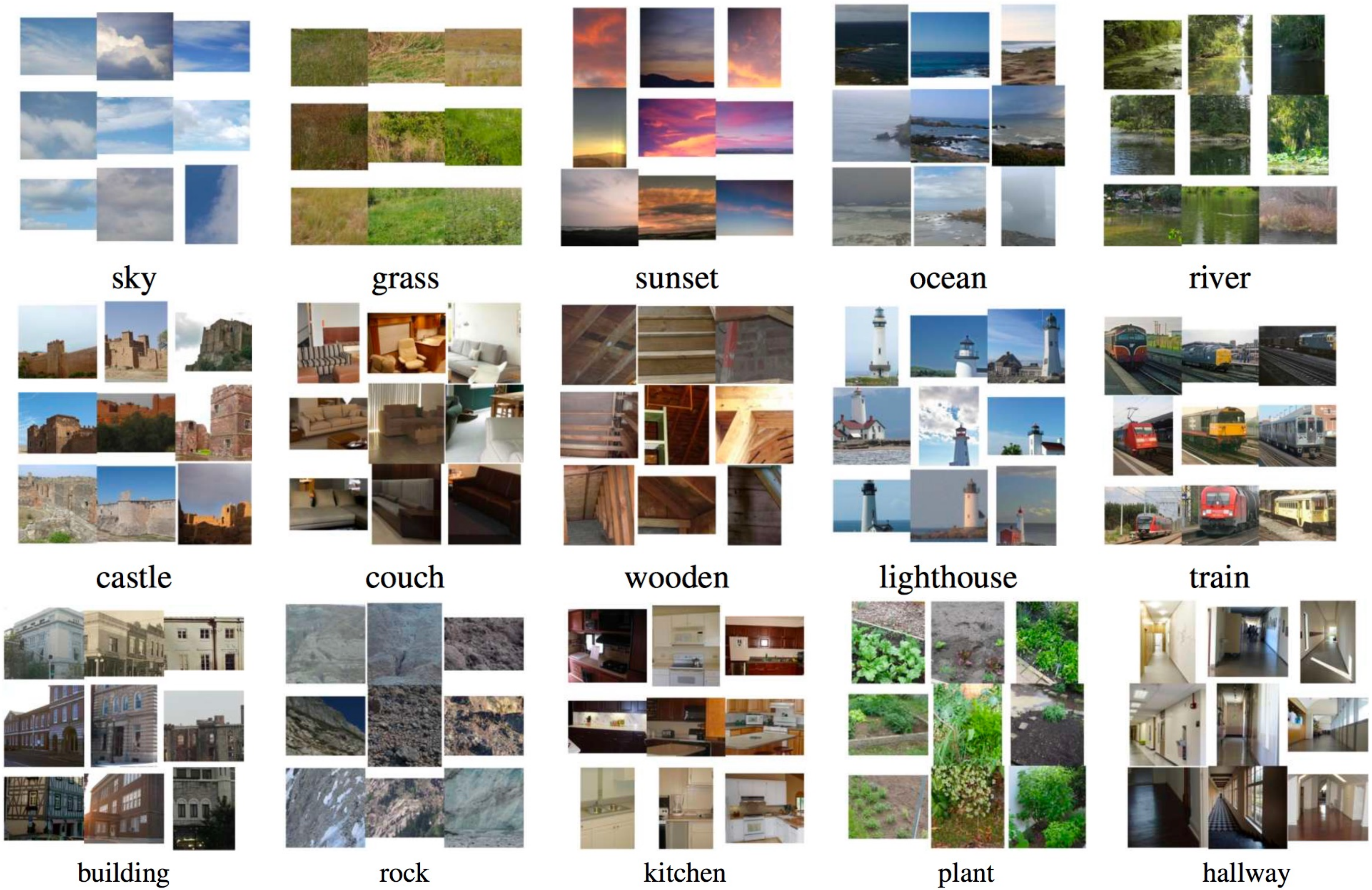


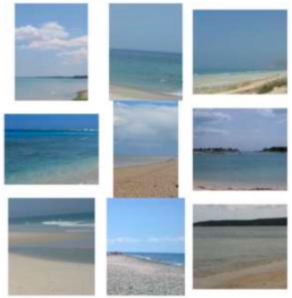




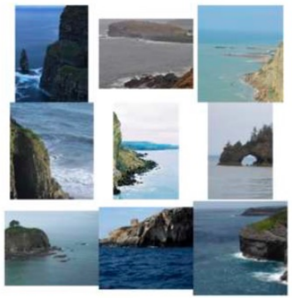
Examples of audio-visual clusters



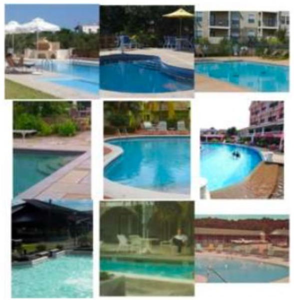

sky
grass
sunset
ocean
river
castle
couch
wooden
lighthouse
train
building
rock
kitchen
plant
hallway



beach



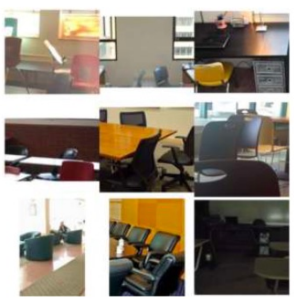
cliff



pool

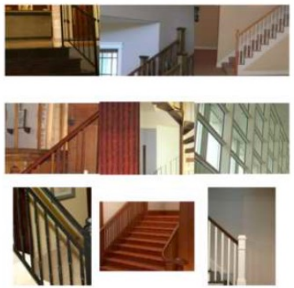
desert

field



chair

table



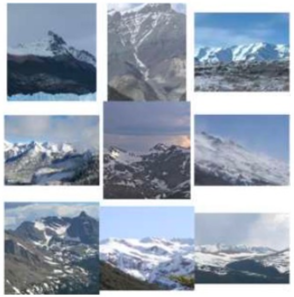
staircase

statue

stone

church

forest



mountain

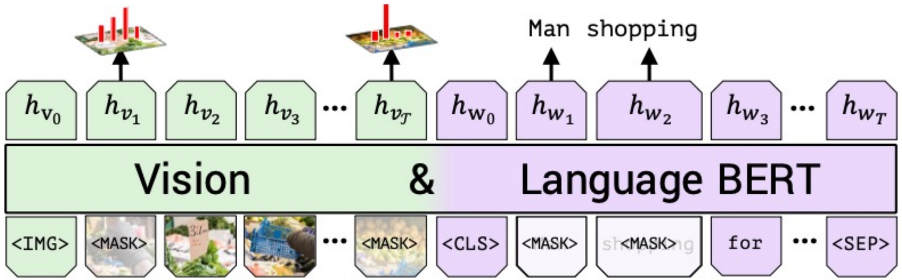
skyscraper

trees

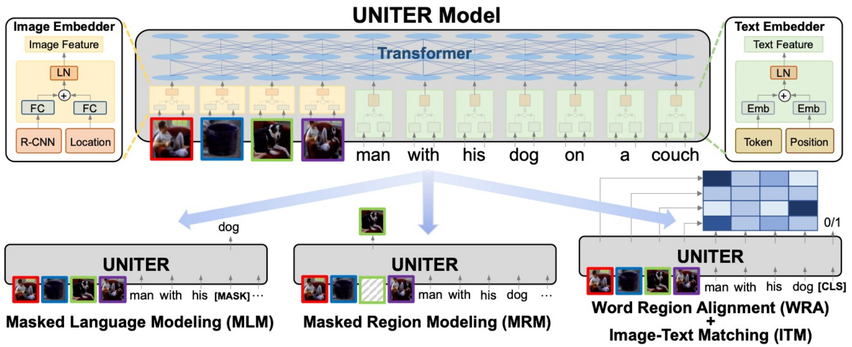

waterfall
windmills
window
city
bridge
flowers
man
wall
archway
baseball
boat
shelves
cockpit
girl
children

Talk Outline

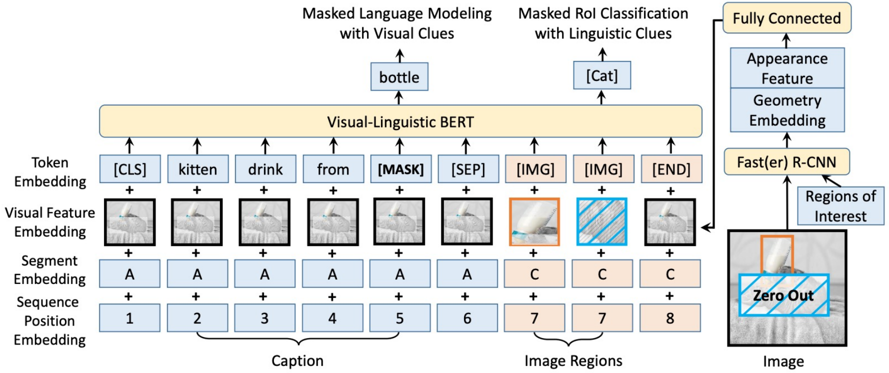
1. Learning representations of speech with visual grounding [Harwath, Torralba, and Glass, NeurIPS 2016], [Harwath and Glass, ACL 2017], [Harwath et al., ECCV 2018]
2. Hybridizing dual-encoders and cross-modal attention models for visually grounding speech [Peng and Harwath, ICASSP 2022]
3. Emergent Word Discovery with Visually-Grounded HuBERT [Peng and Harwath, Interspeech 2022]
4. Learning audio-visual representations of instructional videos in the wild [Rouditchenko et al., Interspeech 2021]
5. Learning to generate spoken image descriptions without text [Hsu, Harwath, Miller, Song, and Glass, ACL 2021]



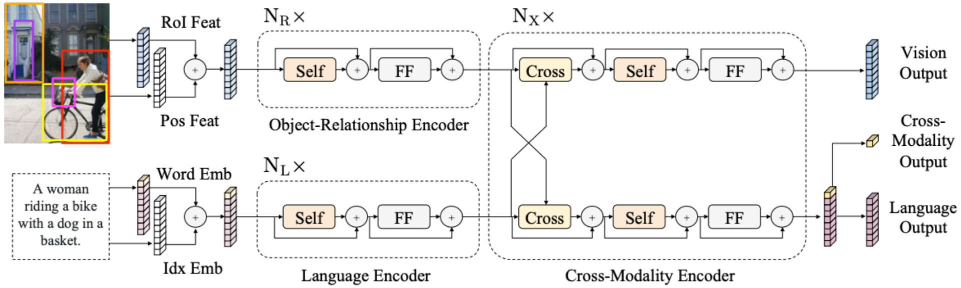
Lu et al., "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," NeurIPS 2019



Chen et al., "UNITER: UNiversal Image-TExt Representation Learning," ECCV 2020

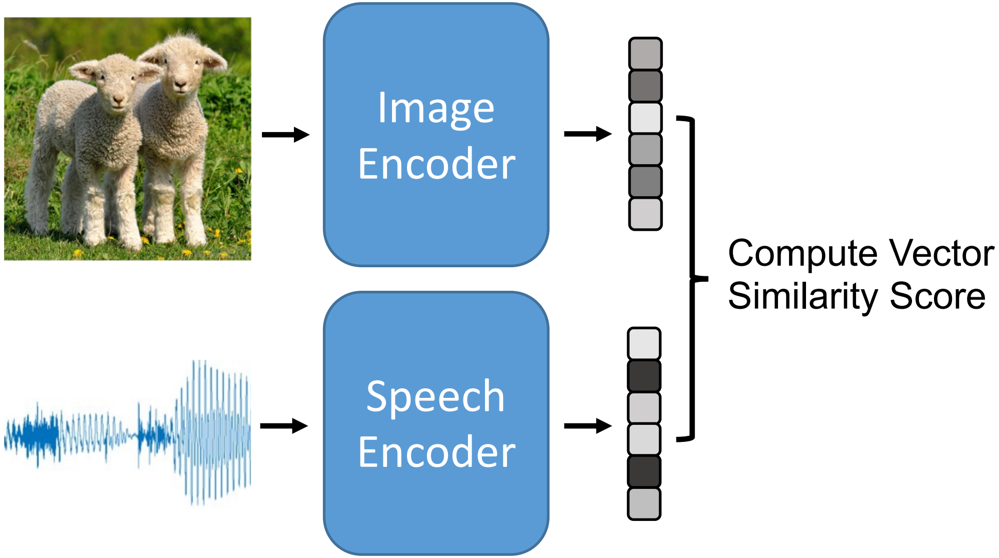


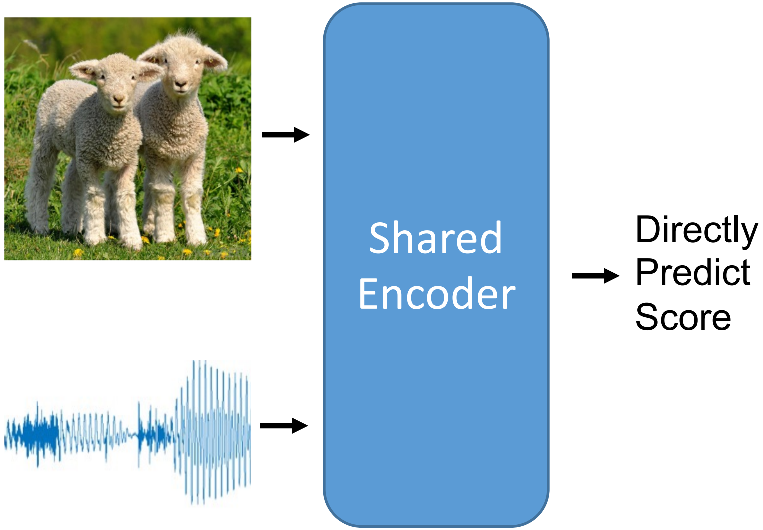
Su et al., "VL-BERT: Pre-Training of Generic Visual-Linguistic Representations," ICLR 2020



Tan and Bansal," LXMERT: Learning Cross-Modality Encoder Representations from Transformers," EMNLP 2019

Dual-Encoders vs. Cross-Attention





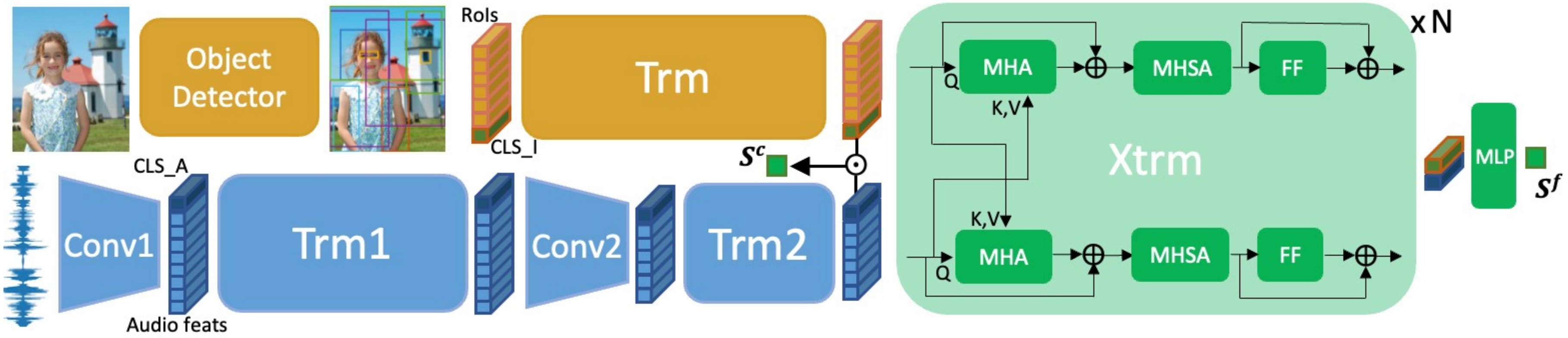
Pros: Simple, lightweight, fast retrieval/scoring if you pre-compute embeddings

Cons: Less powerful at modeling cross-modal interactions

Pros: More powerful at modeling cross-modal interactions

Cons: More expensive to train, can't pre-compute embeddings so retrieval is slower

Fast-Slow Transformer for Visually Grounding Speech (FaST-VGS)

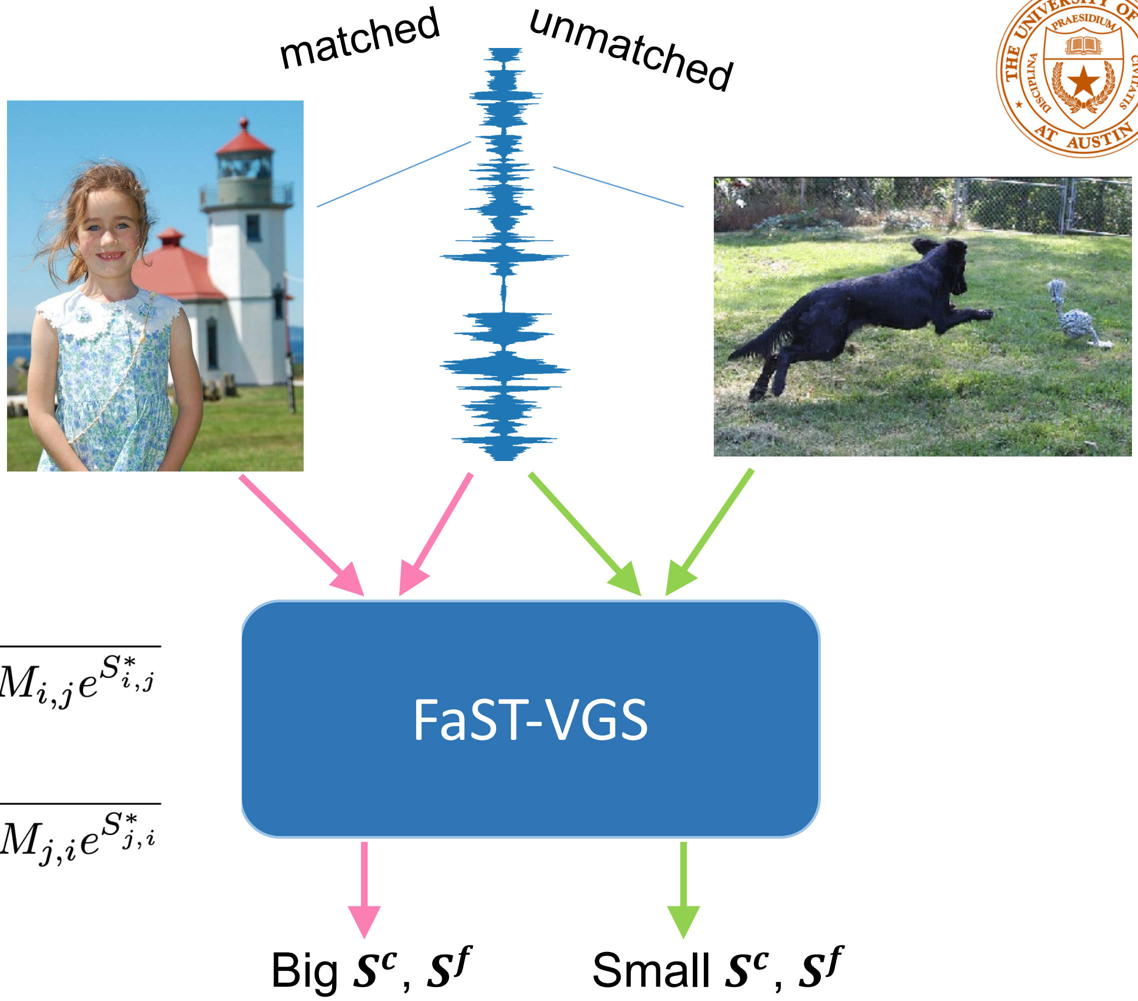


Puyuan Peng and David Harwath. "Fast-slow transformer for visually grounding speech." ICASSP 2022

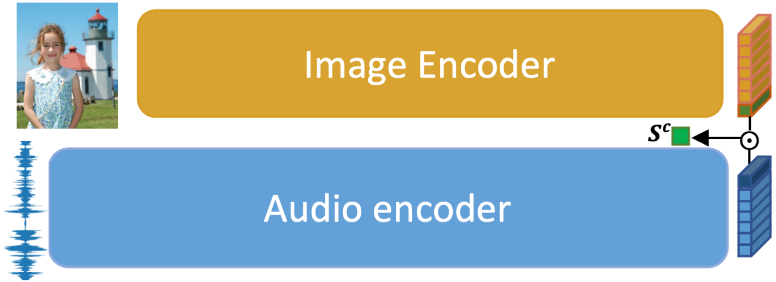
Training

$$\mathcal{L}^*_{A\to I} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{S^*_{i,i}-\delta}}{e^{S^*_{i,i}-\delta}+\sum_{j=1}^{B}M_{i,j}e^{S^*_{i,j}}}$$

$$\mathcal{L}^*_{I\to A} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{S^*_{i,i}-\delta}}{e^{S^*_{i,i}-\delta}+\sum_{j=1}^{B}M_{j,i}e^{S^*_{j,i}}}$$

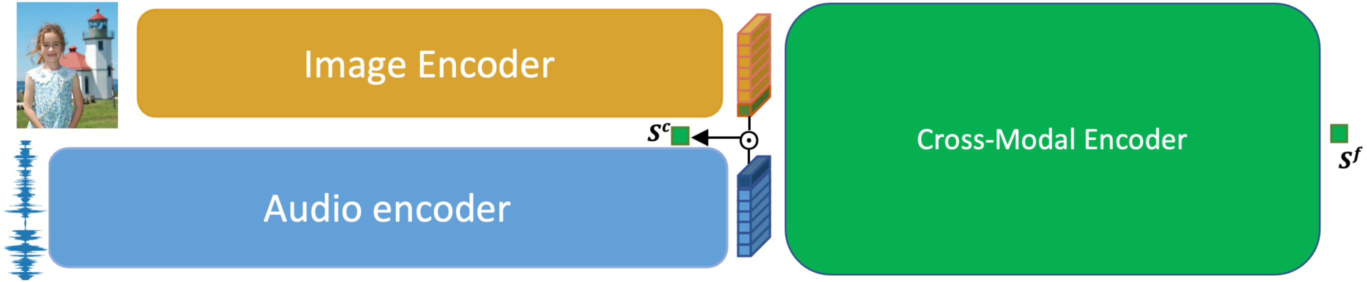


Coarse vs. Fine Retrieval



Coarse

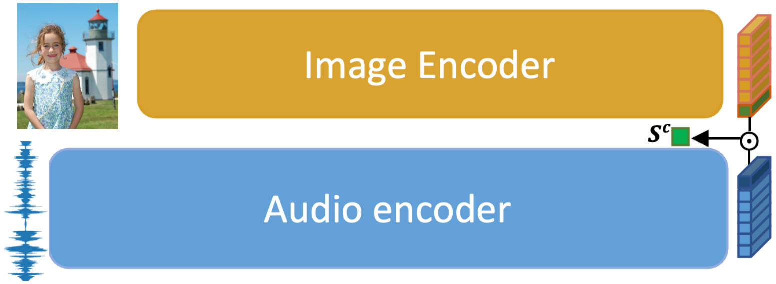
Use scores computed with outputs of dual encoder to perform retrieval



Fine

Use scores computed with outputs of Cross-Modal encoder to perform retrieval

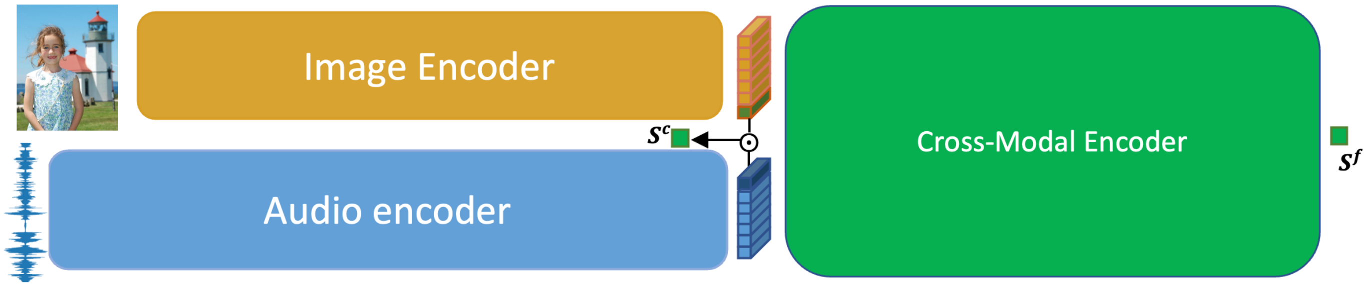
Coarse-To-Fine (CTF) Retrieval



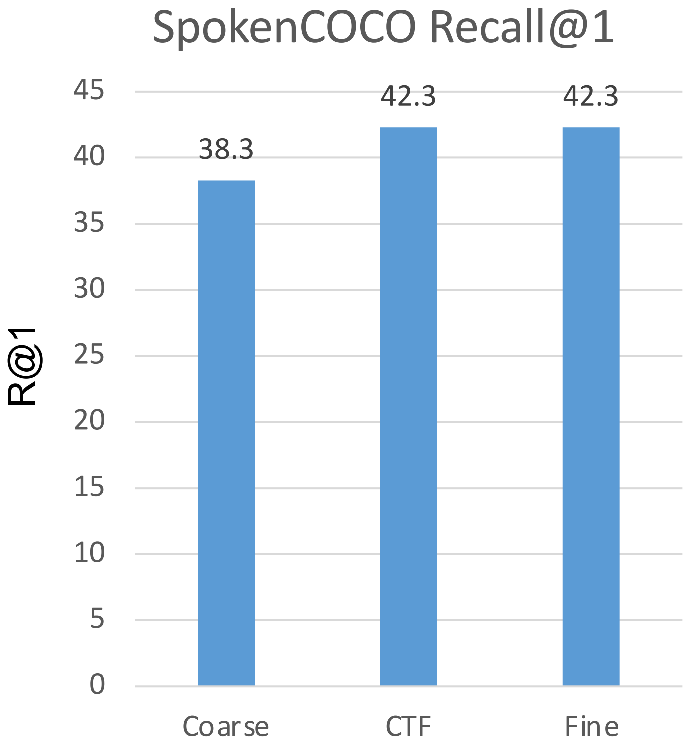
<u>Coarse-To-Fine</u>

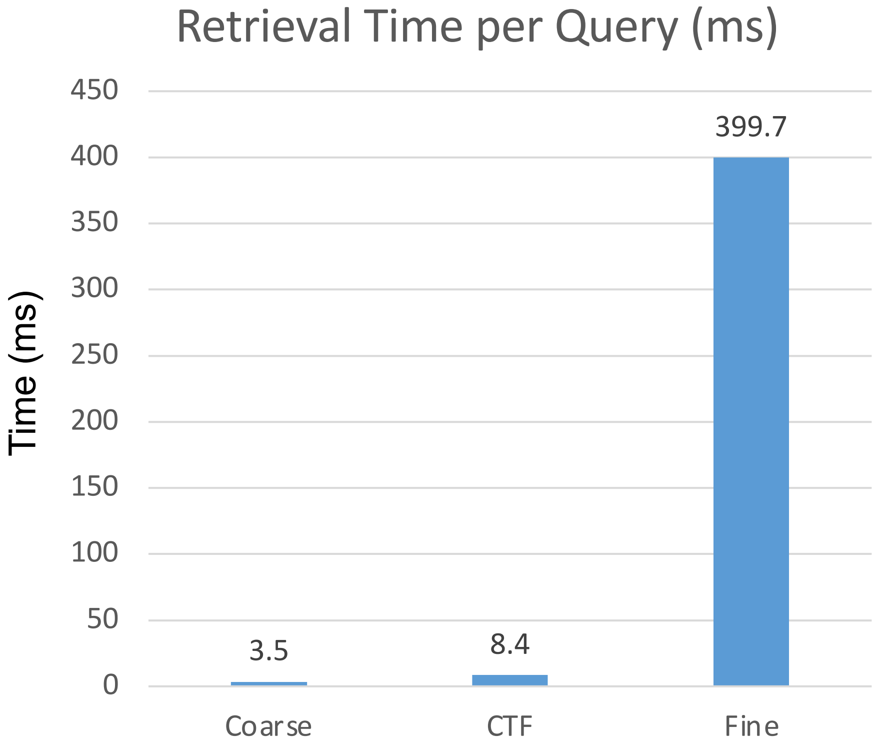
1. Use scores computed with outputs of dual encoder to retrieve the top K items

2. Re-rank the top K items from Step 1 using the Fine retrieval scores

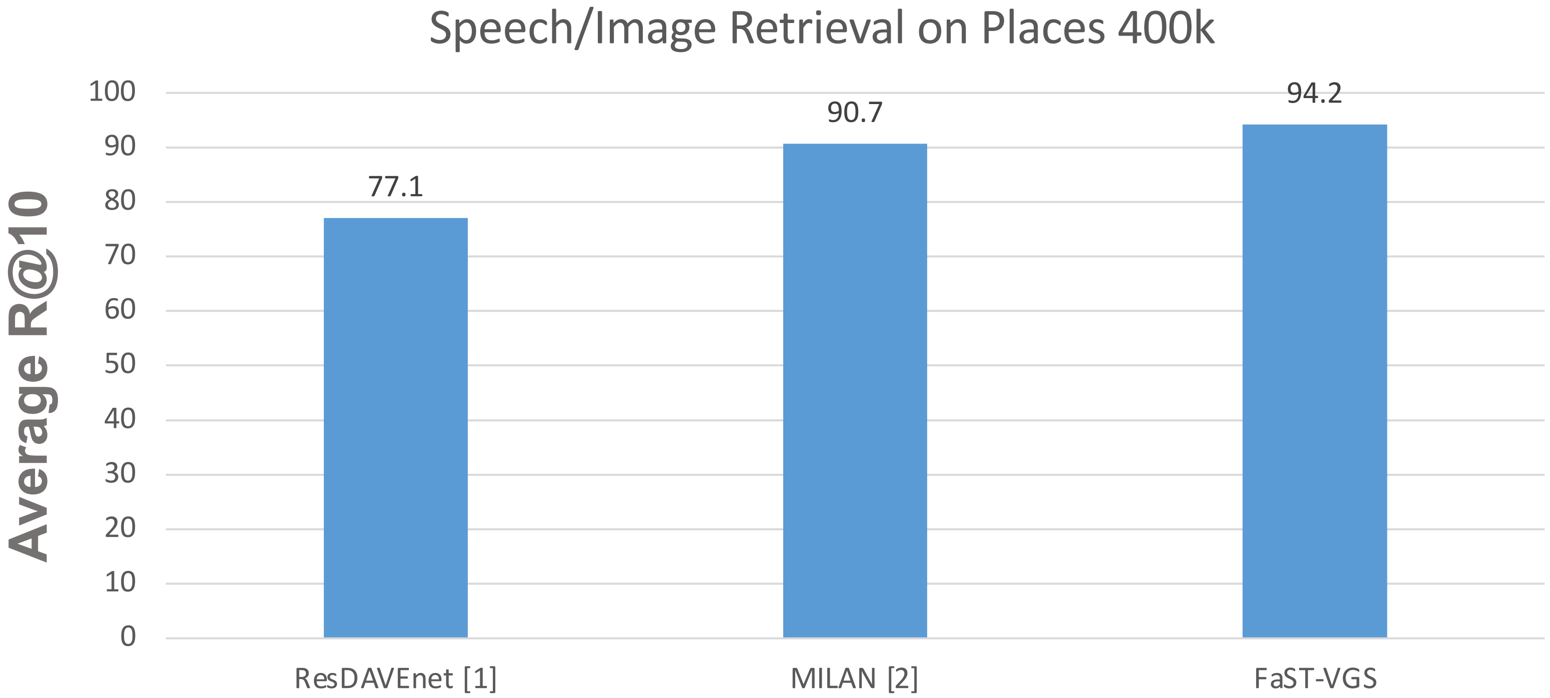


Comparison of Retrieval Methods



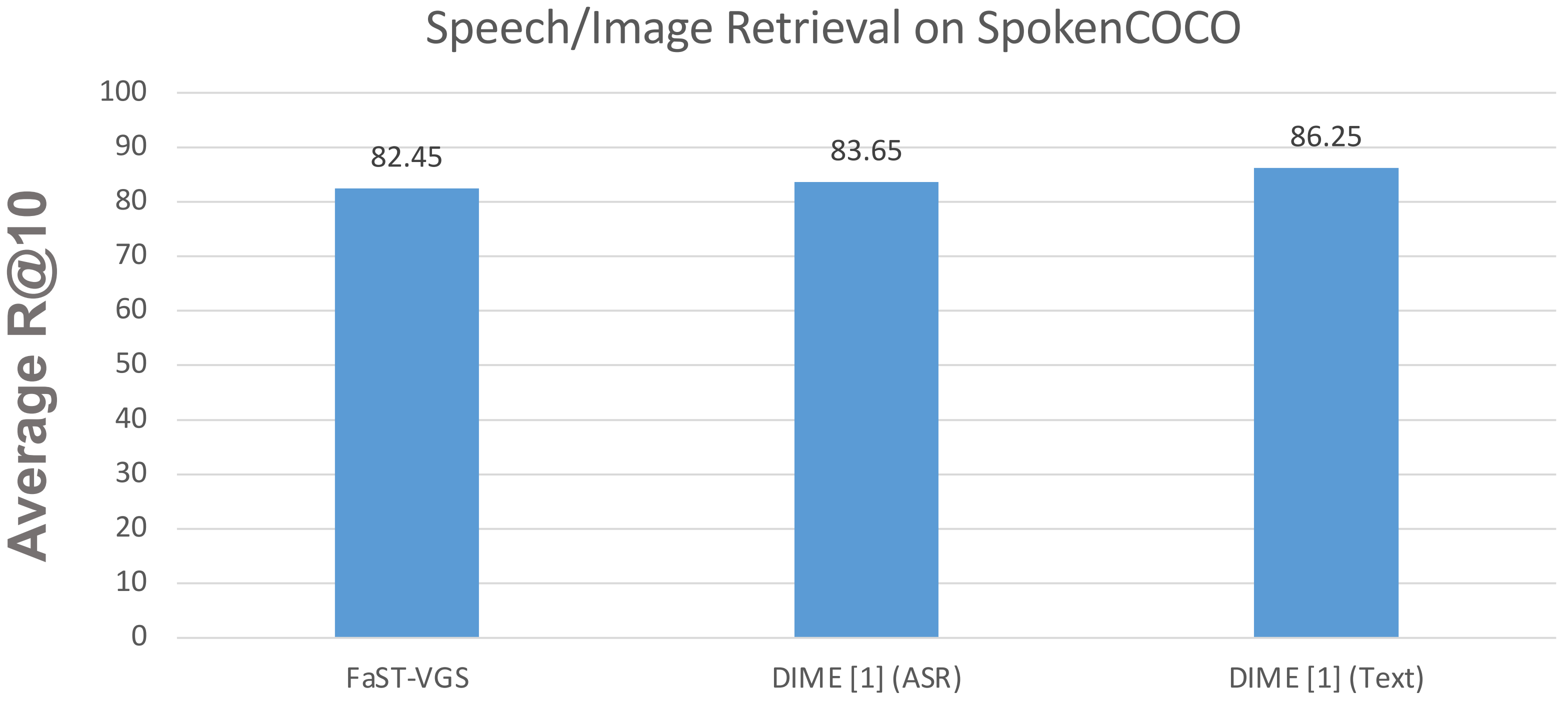


Speech-Image Retrieval on Places400k



[1] D. Harwath, A. Recasens, D. Suris, G. Chuang, A. Torralba, and J. Glass, “Jointly discovering visual objects and spoken words from raw sensory input,” IJCV, 2019.
[2] R. Sanabria, A. Waters, and J. Baldridge, “Talk, don’t write: A study of direct speech-based image retrieval,” Proc. Interspeech, 2021

Comparison to Text-Image Matching Models

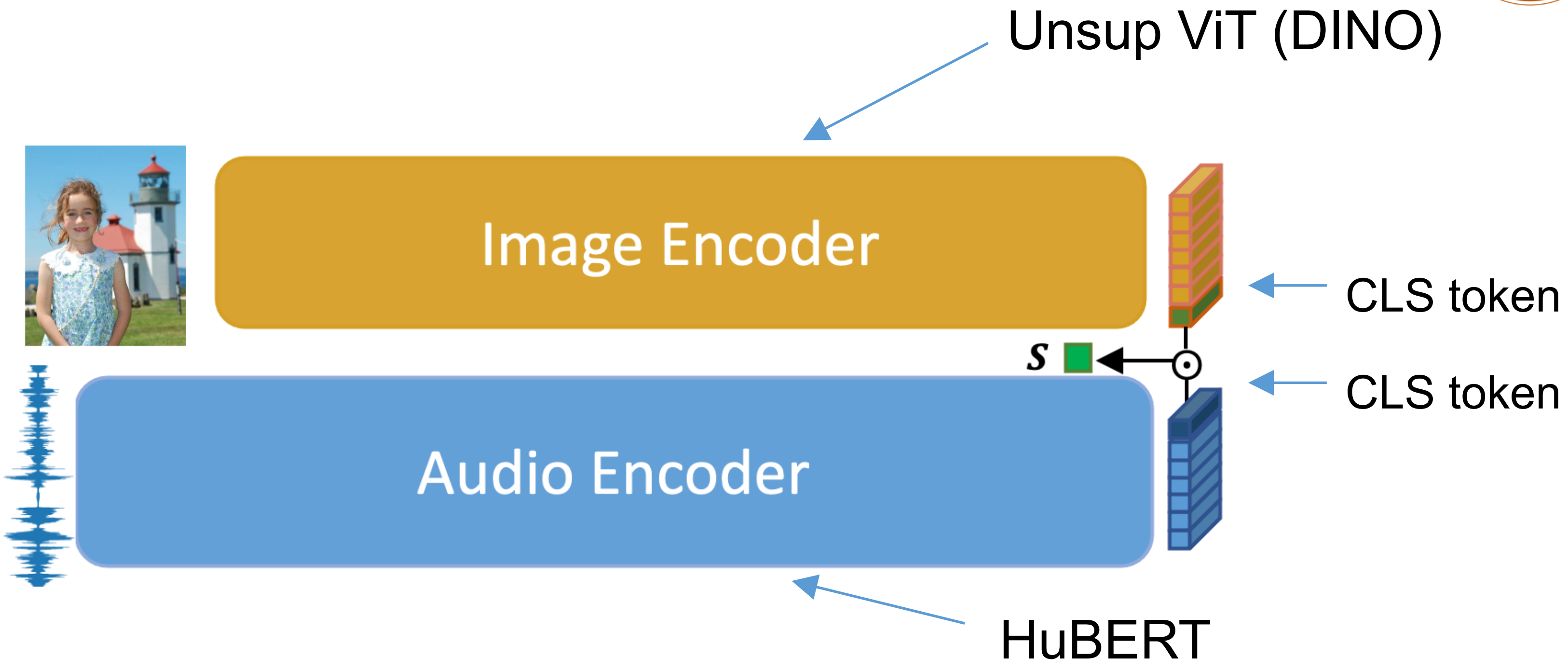


[1] L. Qu, M. Liu, J. Wu, Z. Gao, and L. Nie, "Dynamic modality interaction modeling for image-text retrieval," in ACM SIGIR, 2021

Talk Outline

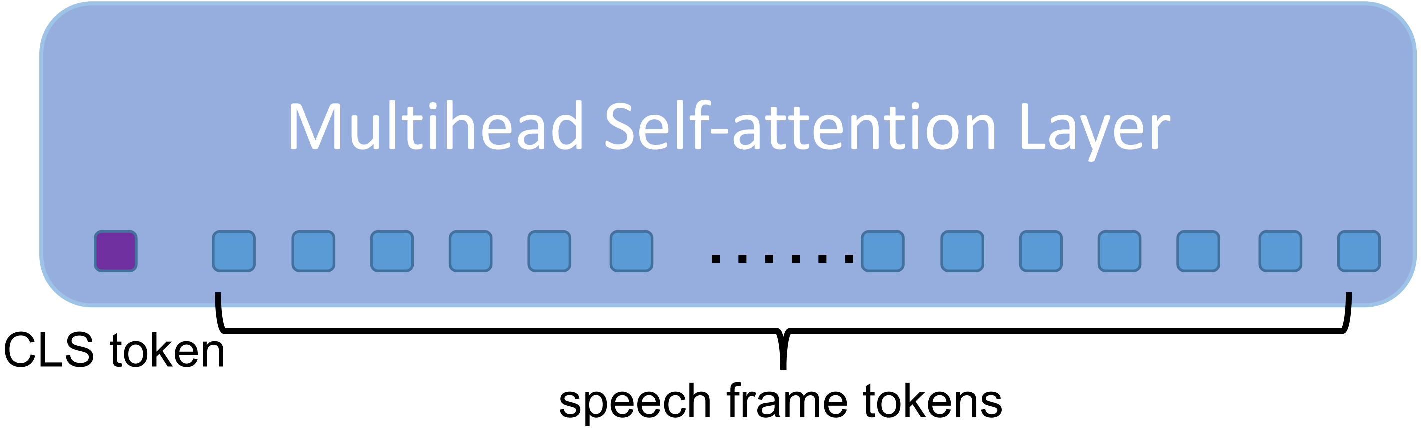
1. Learning representations of speech with visual grounding [Harwath, Torralba, and Glass, NeurIPS 2016], [Harwath and Glass, ACL 2017], [Harwath et al., ECCV 2018]
2. Hybridizing dual-encoders and cross-modal attention models for visually grounding speech [Peng and Harwath, ICASSP 2022]
3. Emergent Word Discovery with Visually-Grounded HuBERT [Peng and Harwath, Interspeech 2022]
4. Learning audio-visual representations of instructional videos in the wild [Rouditchenko et al., Interspeech 2021]
5. Learning to generate spoken image descriptions without text [Hsu, Harwath, Miller, Song, and Glass, ACL 2021]

Visually-Grounded HuBERT (VG-HuBERT)



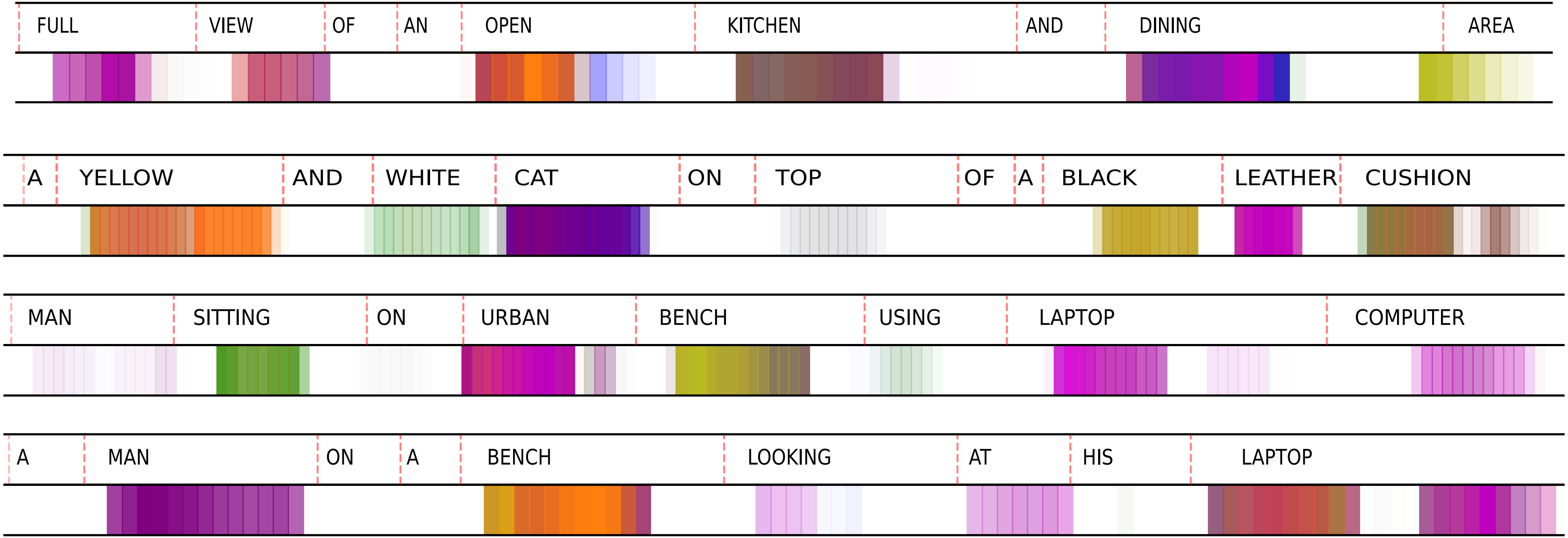
We will examine the self-attention maps of the speech model to see if we can interpret any patterns from them

Visually-Grounded HuBERT (VG-HuBERT)



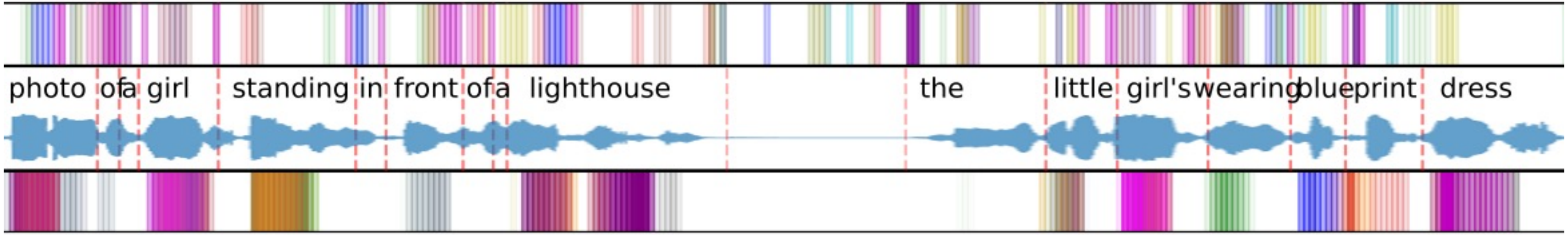
Visualize CLS token's attention to all other speech frame tokens

More Examples



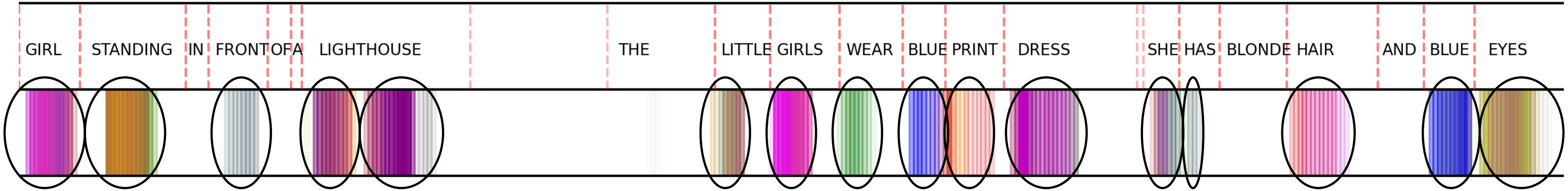
Does HuBERT also discover words?

HuBERT



VG-HuBERT

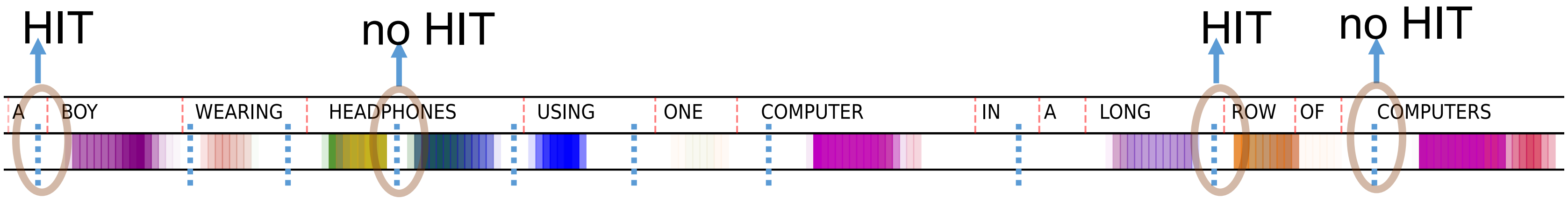
Evaluating Word Discovery



1. Can VG-HuBERT **localize** words?
   71% of words covered by an attention segment on SpokenCOCO test set
2. Can VG-HuBERT **segment** words?
   Word level speech segmentation, measured by precision, recall, F1
3. Can VG-HuBERT **identify** words?
   K-Means clustering on <u>attention segments</u>

Evaluating Word Segmentation

Use mid-points of attention boundaries as predicted word boundaries:



HIT: 1 if a predicted boundary is within ±20ms of a ground truth boundary

Precision = #HIT/#PredictedBoundaries

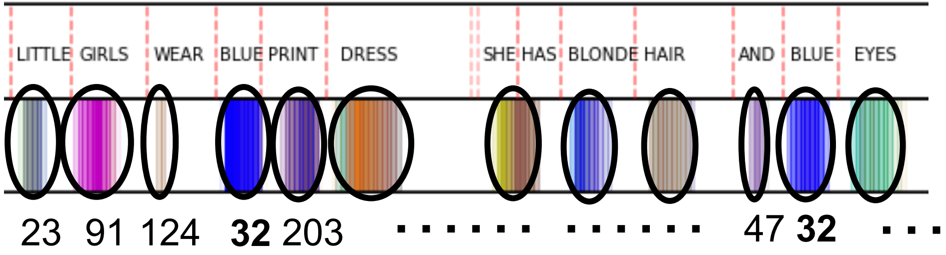
Recall = #HIT/#GTBoundaries

F1 = 2*Precision*Recall / (Precision + Recall)

Segmentation Results on Buckeye

Model	Boundary				Token
	Prec.	Rec.	F_1	R-val.	F_1
Adaptor gram. [45]	15.9	**57.7**	25.0	-139.9	4.4
SylSeg [46]	27.7	28.9	28.3	37.7	19.3
ES-KMeans [6]	30.3	16.6	21.4	39.1	19.2
BES-GMM [5]	31.5	12.4	17.8	37.2	18.6
SCPC [47]	36.9	29.9	33.0	45.6	-
mACPC [10]	<u>42.1</u>	30.3	35.1	<u>47.4</u>	-
DPDP [8]	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**

Evaluating Word Clustering



1. Run KMeans on the mean-pooled CLS attention segmented features

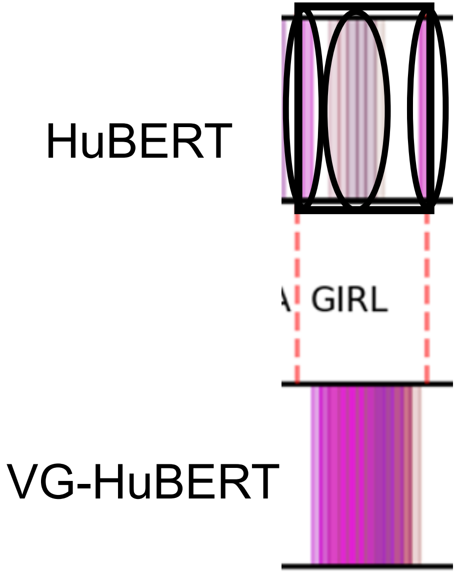
2. Use the KMeans model to assign each segment a cluster number (code)

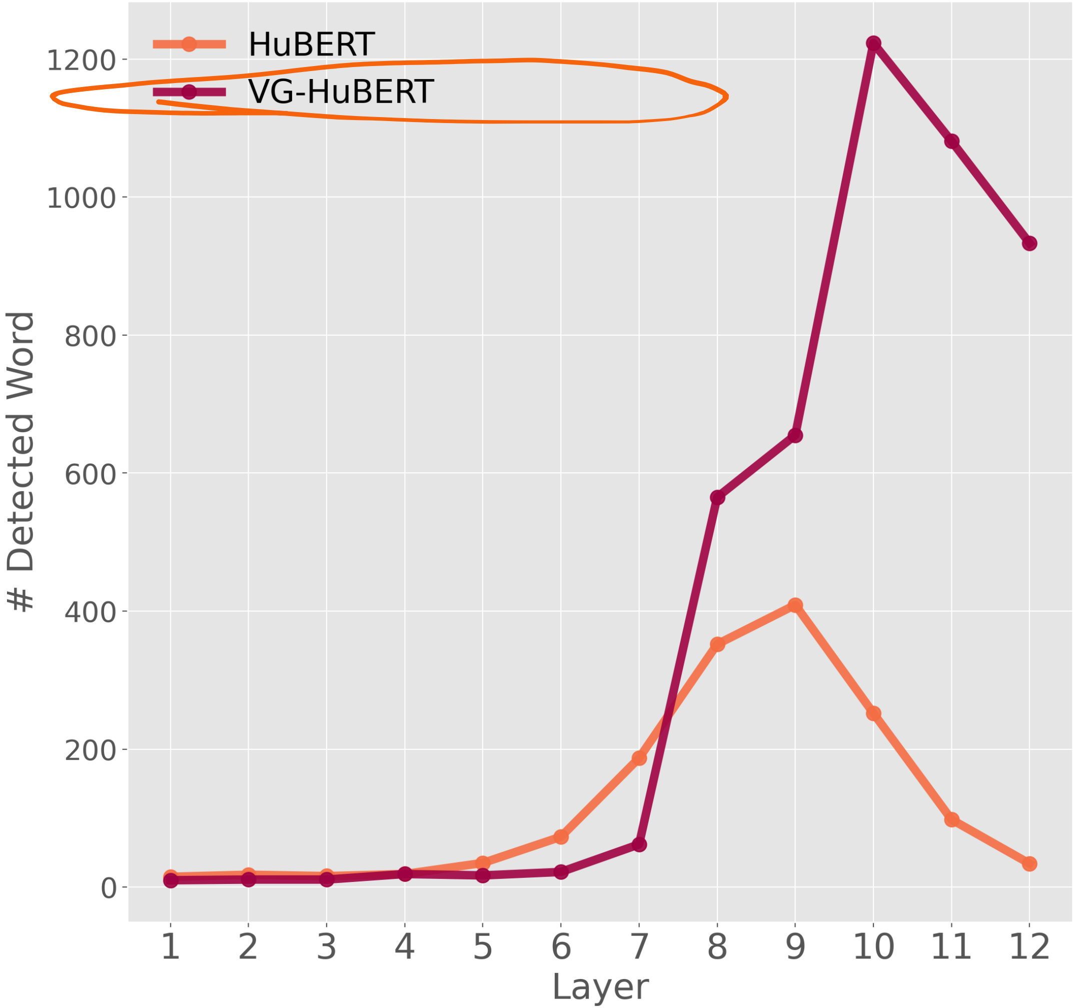
3. Use the assigned codes as word detectors, calculate precision, recall, F1 between code and the word that it overlaps the most with.

Define a word to be detected, if it has F1 > 0.5 with a code

Word detection Results across different layers

SpokenCOCO val set vocab size = 6000





Talk Outline

1. Learning representations of speech with visual grounding [Harwath, Torralba, and Glass, NeurIPS 2016], [Harwath and Glass, ACL 2017], [Harwath et al., ECCV 2018]

2. Hybridizing dual-encoders and cross-modal attention models for visually grounding speech [Peng and Harwath, ICASSP 2022]

3. Emergent Word Discovery with Visually-Grounded HuBERT [Peng and Harwath, Interspeech 2022]

4. Learning audio-visual representations of instructional videos in the wild [Rouditchenko et al., Interspeech 2021]

5. Learning to generate spoken image descriptions without text [Hsu, Harwath, Miller, Song, and Glass, ACL 2021]

Learning language by watching TV

- So far, all of the models I've shown utilize still-frame images that were described by human speakers
- Video and multimedia content on the internet has exploded in volume (30k hours of video uploaded to YouTube every hour)
- Can we leverage this "freely available" data to do cross-modal learning?

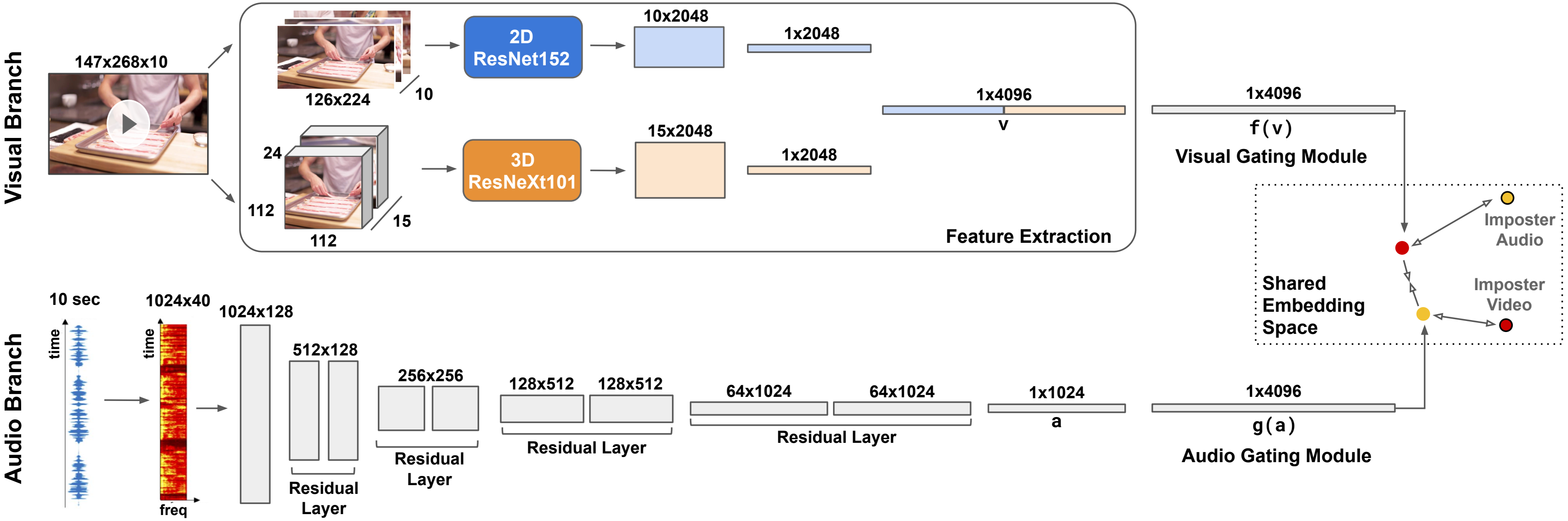
HowTo100M

Meich et al., ICCV 2019

1.2M instructional videos from YouTube (130,000 hours)
23,000 different activities

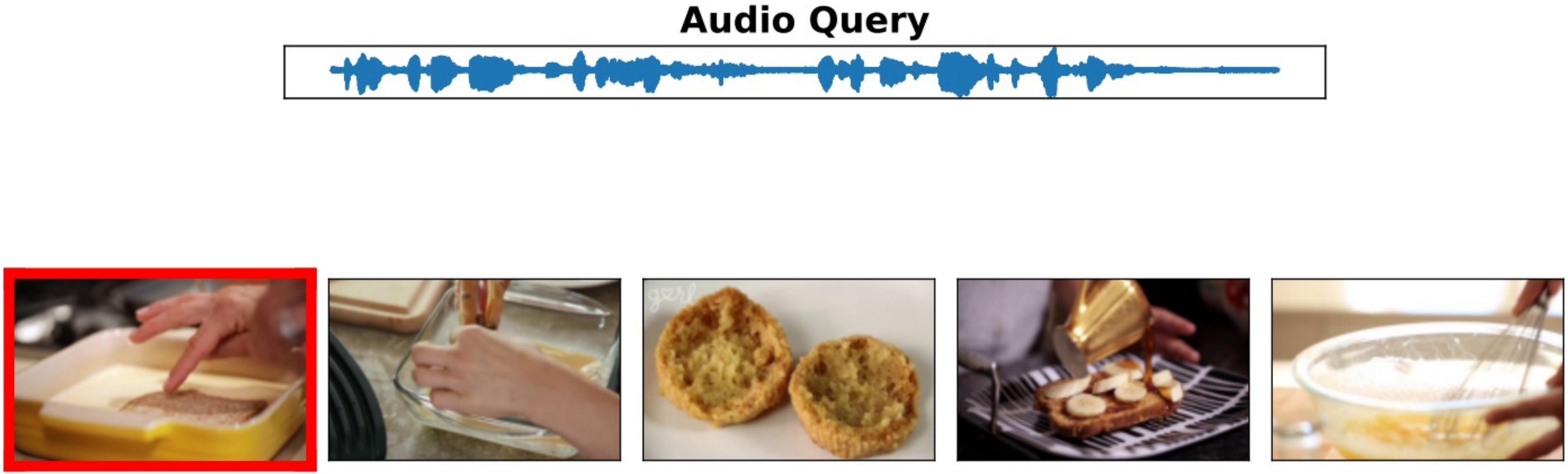
AVLnet Model [Rouditchenko et al., Interspeech 2021]

- Randomly sample 10-second clips from videos
- Treat the co-occurring audio and visual streams as matched pairs
- Train with contrastive objective by sampling mismatched audio/visual streams from other clips in the same minibatch



Retrieval-Based Evaluation

Task: Given the audio stream of a video clip, correctly match it with its corresponding video stream

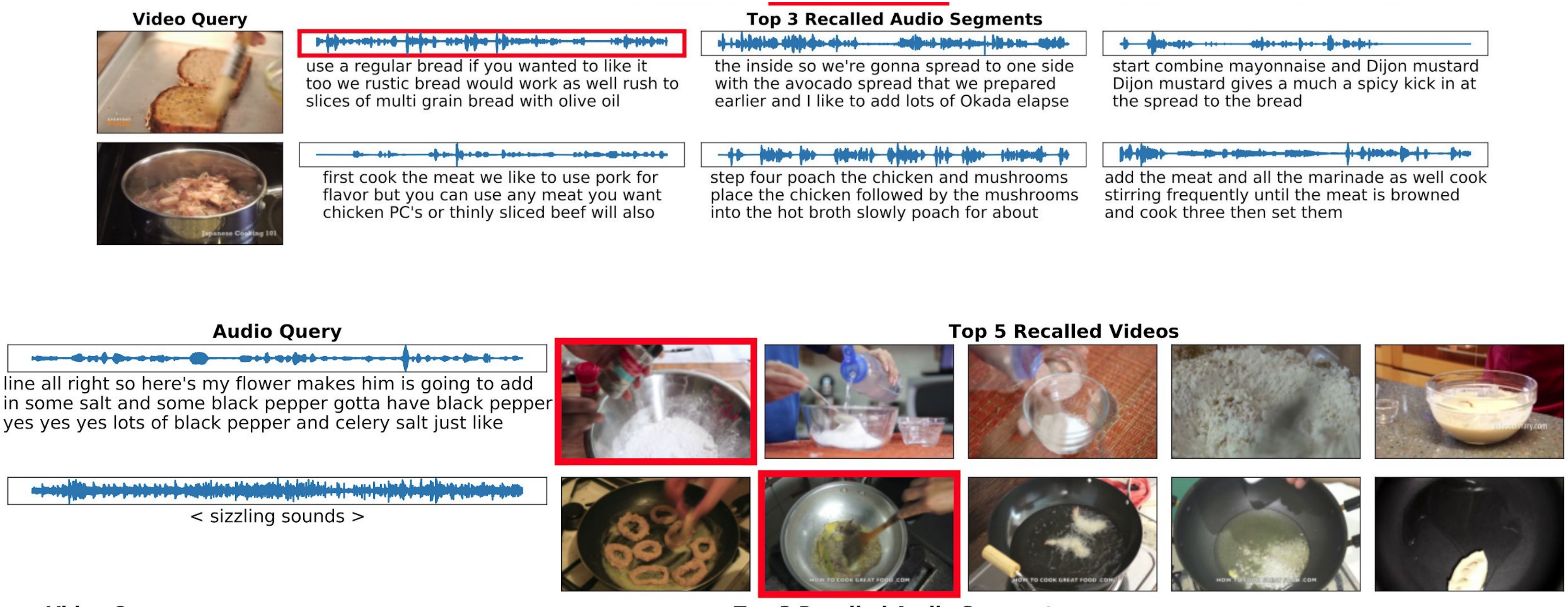


Retrieval Results

(a) Video clip retrieval (A→V).

	Method (A→V)	YouCook2 R@1	YouCook2 R@5	YouCook2 R@10	CrossTask R@1	CrossTask R@5	CrossTask R@10	MSR-VTT R@1	MSR-VTT R@5	MSR-VTT R@10
	Random	0.03	0.15	0.3	0.04	0.18	0.35	0.1	0.5	1.0
No pretraining	(1) Boggust et al. [59]	0.5	2.1	3.4	0.4	1.9	3.7	1.0	3.8	7.1
	(1) Arandjelović et al. [11]	0.3	1.9	3.3	0.4	**2.5**	4.1	**1.3**	4.3	8.2
	(1) **AVLnet**	**0.7**	**2.3**	**3.9**	**0.7**	2.4	**4.6**	0.9	**5.0**	**9.0**
HowTo100M Pretrained (Zero Shot)	(2) Boggust et al. [59]	6.8	22.4	31.8	5.5	18.7	28.3	7.6	21.1	28.3
	(2) Arandjelović et al. [11]	13.6	31.7	41.8	7.3	19.5	27.2	12.6	26.3	33.7
	(2) **AVLnet**	**27.4**	**51.6**	**61.5**	**11.9**	**29.4**	**37.9**	**17.8**	**35.5**	**43.6**
HowTo100M Pretrained (Fine-Tuned)	(3) Boggust et al. [59]	8.5	26.9	38.5	6.6	20.8	31.2	10.3	27.6	35.9
	(3) Arandjelović et al. [11]	17.4	39.7	51.5	9.5	25.8	36.6	16.2	32.2	42.9
	(3) **AVLnet**	**30.7**	**57.7**	**67.4**	**13.8**	**34.5**	**44.8**	**20.1**	**40.0**	**49.6**

Qualitative Results



Talk Outline

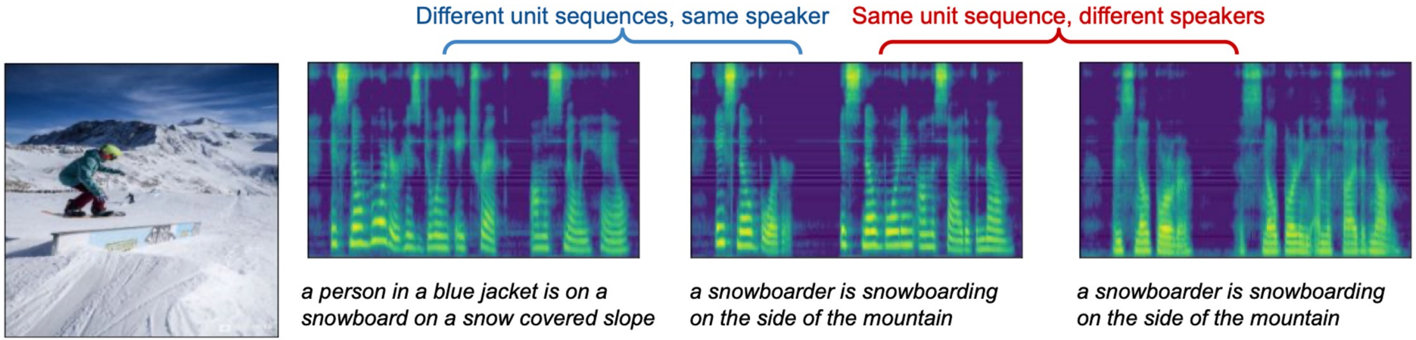
1. Learning representations of speech with visual grounding [Harwath, Torralba, and Glass, NeurIPS 2016], [Harwath and Glass, ACL 2017], [Harwath et al., ECCV 2018]
2. Hybridizing dual-encoders and cross-modal attention models for visually grounding speech [Peng and Harwath, ICASSP 2022]
3. Emergent Word Discovery with Visually-Grounded HuBERT [Peng and Harwath, Interspeech 2022]
4. Learning audio-visual representations of instructional videos in the wild [Rouditchenko et al., Interspeech 2021]
5. Learning to generate spoken image descriptions without text [Hsu, Harwath, Miller, Song, and Glass, ACL 2021]

Motivation

Goal: build a system capable of generating fluent speech describing an image *without using any text supervision*

Why:

1. Humans can learn to speak before they learn to read and write
2. Most text-free speech studies focus on inference but not generation
3. Image-to-text is studied extensively, providing reusable metrics



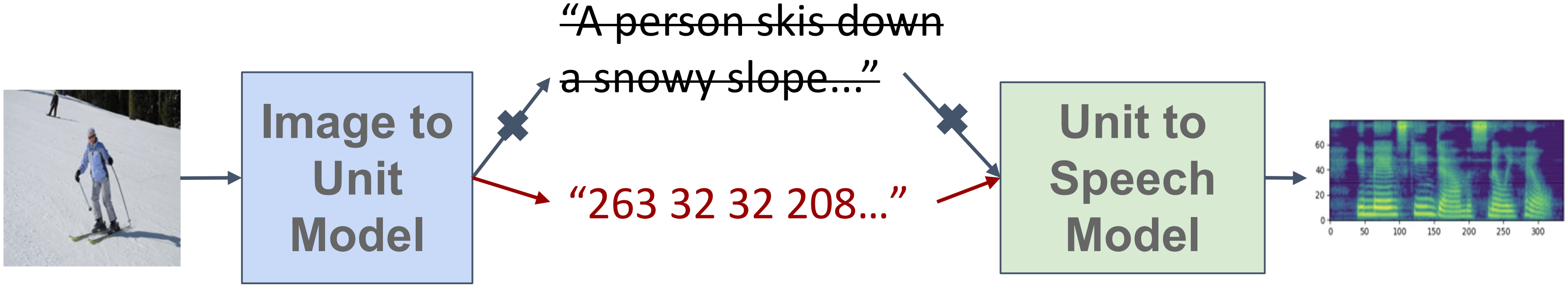
Discovered Units as A Drop-In Replacement for Text

Even for languages without a commonly used writing system, there are still *inventories of words and phonemes*

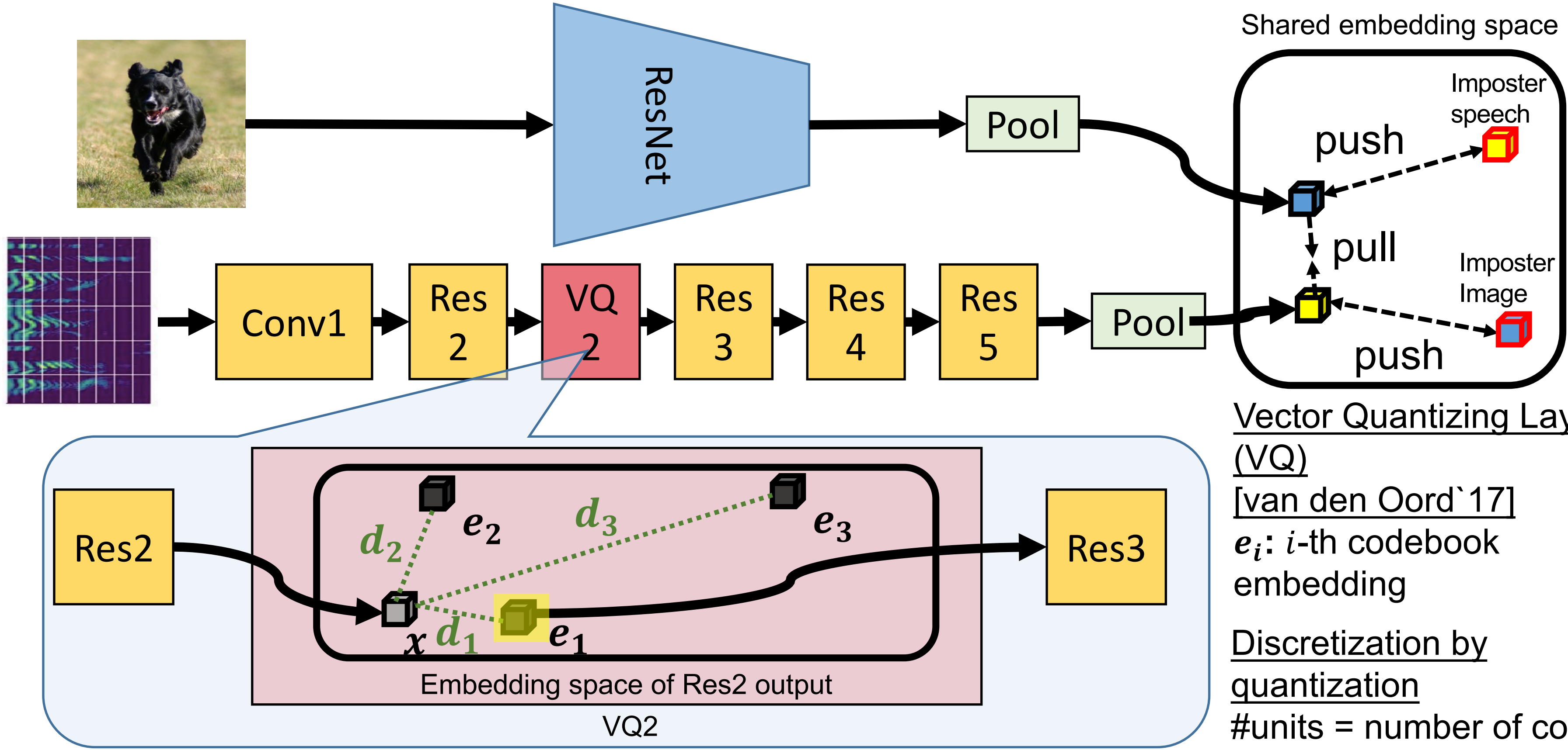
- Leverage automatically discovered units to replace text!

Benefit of the pipeline system

1. Exploits the development from text-based systems
2. Use separate data to train each module



ResDAVEnet-VQ Speech Units



Vector Quantizing Layer (VQ) [van den Oord`17]

e_i: i-th codebook embedding

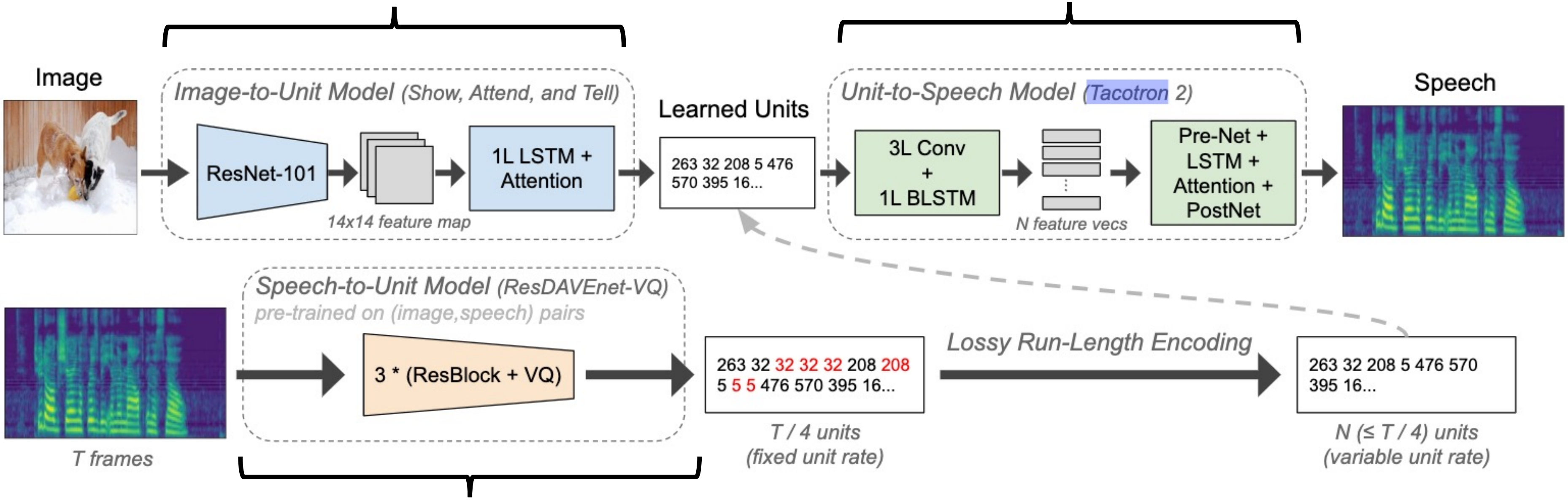
Discretization by quantization

#units = number of codes

Detailed Model Diagram

Image captioning model trained to predict latent units instead of words

Text-to-speech model trained to take as input latent units instead of phones/characters/words



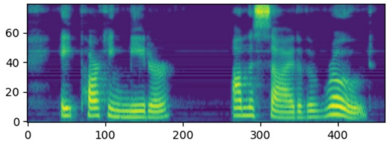
Pre-trained ResDAVEnet-VQ model

Beam Search is Successful Only with Robust Units + RLE

Symbol	Captioned Generated with Beam Search (beam size=5)
VQ3	263 32 208 5 336 100 717 803 256 803 815 144 120 144 654 936 48 417 272 417 362 766 825 284 614 156 341 135 769 5 208 32 208 5 336 815 144 815 494 181 467 417 870 395 683 141 250 543 820 587 181 913 1013 467 5 208 32 208 5 467 360 606 360 801 1009 398 847 89 100 869 254 1003 442 42 791 417 272 141 766 362 614 156 341 135 769 5 208 32
VQ2	71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791...
WVQ	181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232...
VQ3 \ RLE	263 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32...

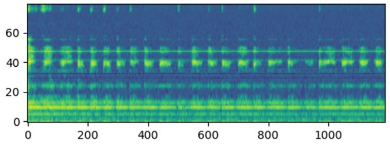
Repeating bigrams

Repeating unigrams



VQ3 (with RLE)

ASR: a parking meter on the side of the road



WVQ (with RLE)

ASR: a esna ey area of a ey area

unit-per-second	Char < WVQ = VQ3 < VQ2
quality (ABX)	Char > VQ2 > VQ3 >> WVQ
duration info	Char < RLE < Plain

Quantitative Results

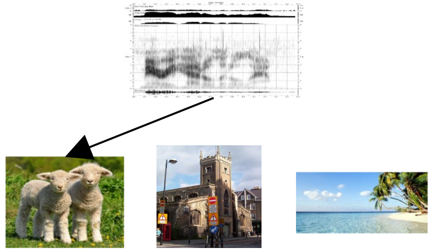
symbol	Greedy / Beam-Search (SAT Model)				
	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
word	0.287 / 0.315	0.247 / 0.253	0.524 / 0.533	0.939 / 0.984	0.180 / 0.185
char	0.238 / 0.289	0.230 / 0.239	0.495 / 0.512	0.783 / 0.879	0.164 / 0.172
VQ3	0.133 / 0.186	0.162 / 0.186	0.413 / 0.446	0.435 / 0.584	0.111 / 0.127
VQ2	0.068 / 0.073	0.138 / 0.126	0.343 / 0.345	0.262 / 0.224	0.084 / 0.065
WVQ	0.010 / 0.009	0.069 / 0.069	0.286 / 0.285	0.009 / 0.009	0.011 / 0.011
VQ3 \ RLE	0.000 / 0.000	0.002 / 0.002	0.001 / 0.001	0.000 / 0.000	0.001 / 0.001

- Evaluation method: use ASR on generated speech, then compare to ground truth text captions
- VQ3 + RLE is still behind word/char, but with non-trivial scores


E2
C

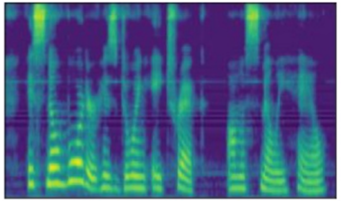

Sanitarium
Granola
BERRI
Orange

What I’ve shown today



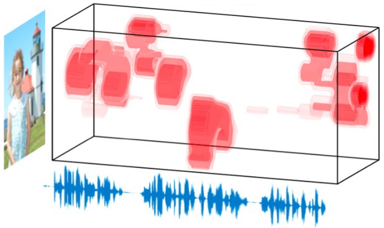
Multimedia retrieval

Discovering Objects

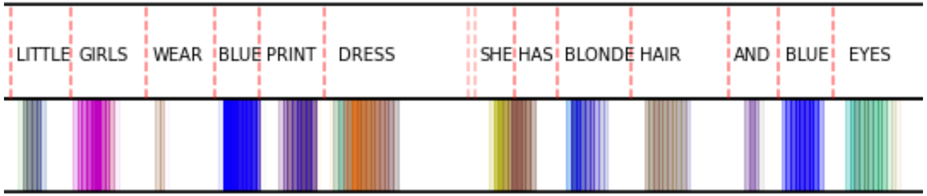


a person in a blue jacket is on a snowboard on a snow covered slop

Textless Image-to-Speech Captioning

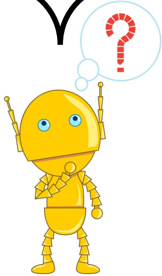


Cross-modal semantics



Discovering Words

Cross-modal, self-supervised learning



Read about our work in more detail at:

https://saltlab.cs.utexas.edu/

And now for a live demo…

Are Different Units/Encoding Equally Suitable?

Three units from two Speech-to-Unit Models

1. ResDAVEnet-VQ (**VQ2** / **VQ3**)
    a. trained for grounding, more robust
    b. VQ3 down-samples more -> lower unit rate
2. WaveNet-VQ (**WVQ**)
    a. trained for reconstruction

unit-per-second	Char < WVQ = VQ3 < VQ2
quality (ABX)	Char > VQ2 > VQ3 >> WVQ
duration info	Char < RLE < Plain

Two encoding methods

1. **Plain**: fixed unit-rate
2. **Run-Length Encoding (RLE)**: remove repetition
    a. [1, 1, 1, 2, 2] -> [1, 2]

How to Evaluate Image-to-Speech Models?

Unit-Based Evaluation

- **Unit-BLEU**: not comparable across units

Text-Based Evaluation

- **Word-BLEU**: adjusted n-gram precision
- **SPICE**: F1 score on semantic scene graph

Text-Free Evaluation

- Recall@N for cross-modal retrieval

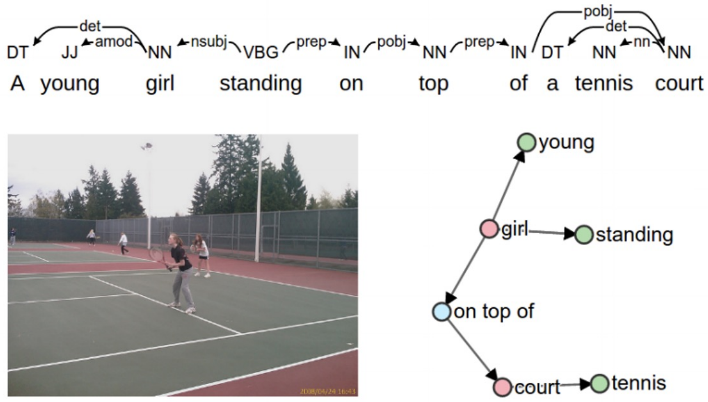


Illustration of scene graphs from Anderson et al. (2016)

Experimental Setup

Data

- **Speech-to-Unit**: PlacesAudio 400k (*non-scripted crowdsourced speech*)
- **Image-to-Unit**: SpokenCOCO (*real scripted speech based on MSCOCO*)
- **Unit-to-Speech**: LJSpeech (*single speaker TTS dataset*)

Model

- **Speech-to-Unit:** ResDAVEnet-VQ [Harwath et al. 2019], WaveNet-VQ [Chorowski et al., 2019]
- **Image-to-Unit**: Show-Attend-Tell [Xu et al., 2016]
- **Unit-to-Speech**: Tacotron2 [Shen et al., 2018] + WaveGlow [Prenger et al., 2018]

Units need to be *robust* to bridge different systems well!

SUPERB: Speech processing Universal PERformance Benchmark

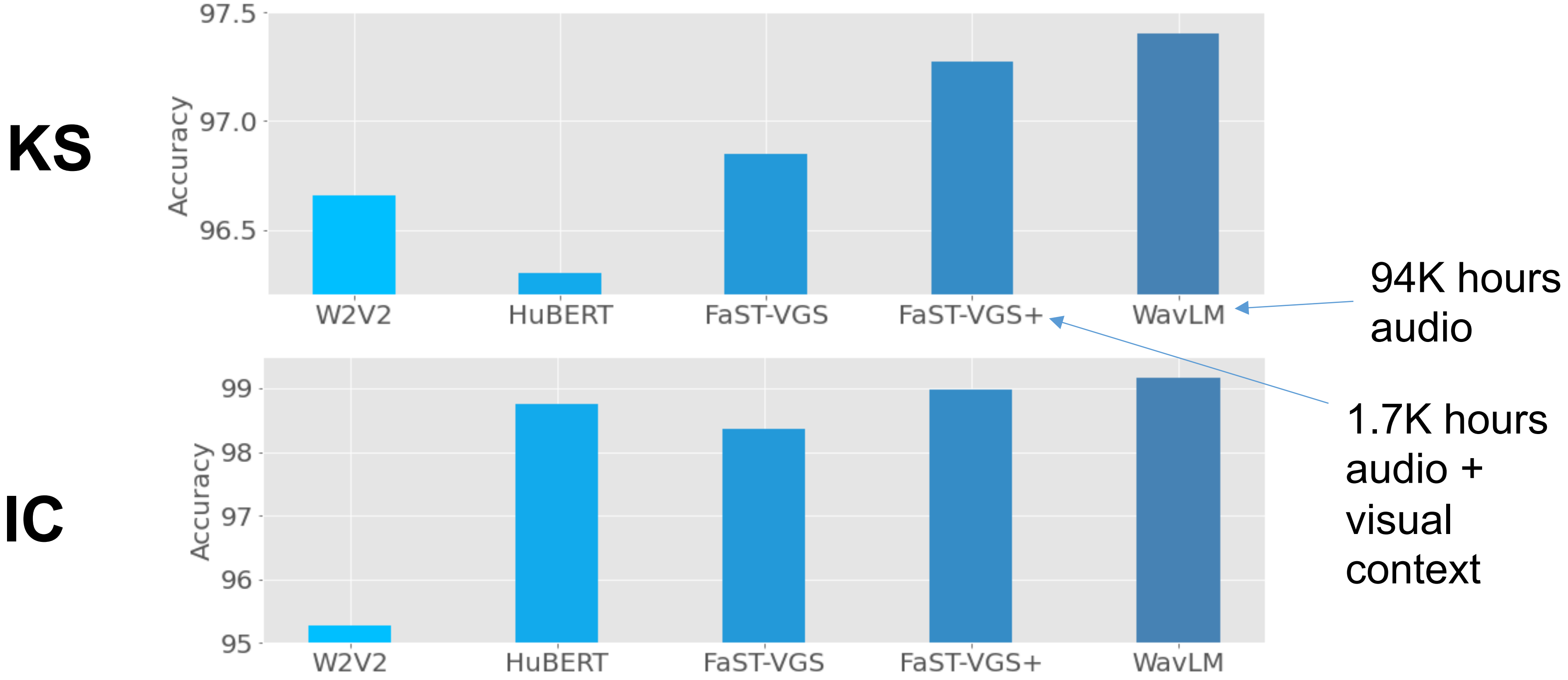
Shu-wen Yang[1], *Po-Han Chi*[1*], *Yung-Sung Chuang*[1*], *Cheng-I Jeff Lai*[2*], *Kushal Lakhotia*[3*], *Yist Y. Lin*[1*], *Andy T. Liu*[1*], *Jiatong Shi*[4*], *Xuankai Chang*[6], *Guan-Ting Lin*[1], *Tzu-Hsien Huang*[1], *Wei-Cheng Tseng*[1], *Ko-tik Lee*[1], *Da-Rong Liu*[1], *Zili Huang*[4], *Shuyan Dong*[5†], *Shang-Wen Li*[5†], *Shinji Watanabe*[6], *Abdelrahman Mohamed*[3], *Hung-yi Lee*[1]

[1]National Taiwan University, Taiwan
[2]Massachusetts Institute of Technology, USA
[3]Facebook AI Research, USA
[4]Johns Hopkins University, USA
[5]Amazon AI, USA
[6]Carnegie Mellon University, USA

SUPERB Results

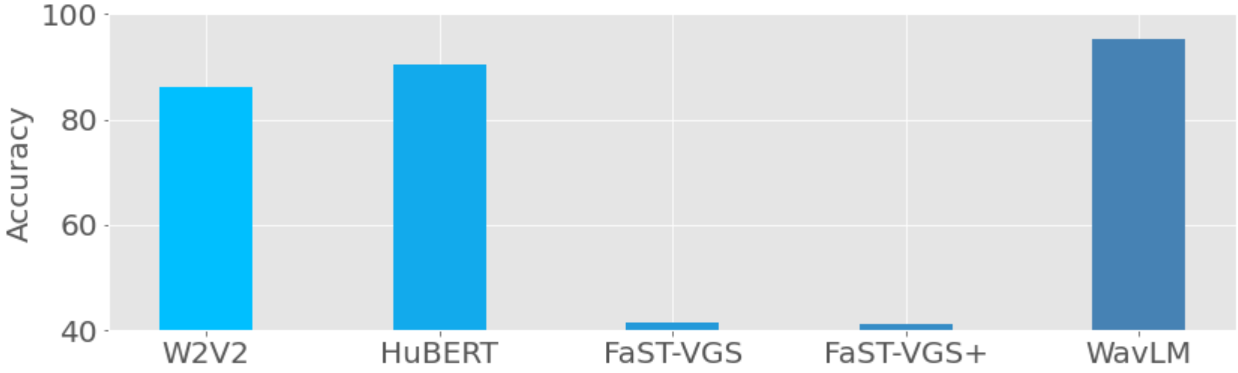
Method	#Params	Data	Speaker			Content					Semantics			ParaL
			SID	ASV	SD	PR	ASR (WER)		KS	QbE	IC	SF		ER
			Acc ↑	EER ↓	DER ↓	PER ↓	w/o ↓	w/LM ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	15.21	8.63	0.0058	9.10	69.64	52.94	35.39
PASE+	7.83M	LS50	37.99	11.61	8.68	58.87	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**

FaST-VGS+ performs well on keyword spotting and intent classification



FaST-VGS+ performs decently on ASR but poorly on Speaker ID

SID



ASR