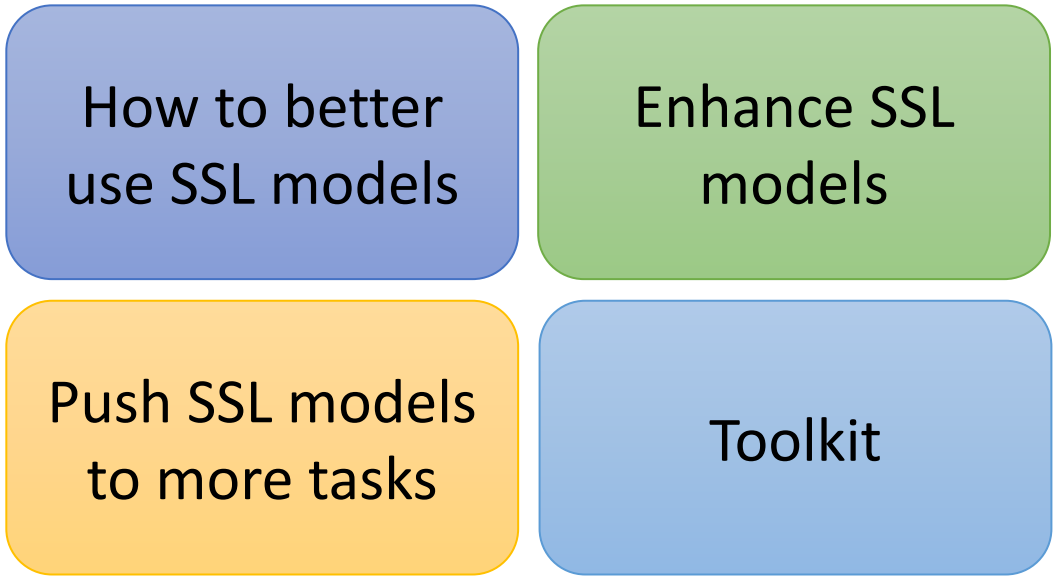
Leveraging Pre-training Models for Speech Processing

Research Group @ JSALT 2022

Goal


How to better use SSL models
Enhance SSL models
Push SSL models to more tasks
Toolkit

Goal

- More efficient
- Better generalization
- Visual enhanced

How to better use SSL models

Enhance SSL models

Push SSL models to more tasks

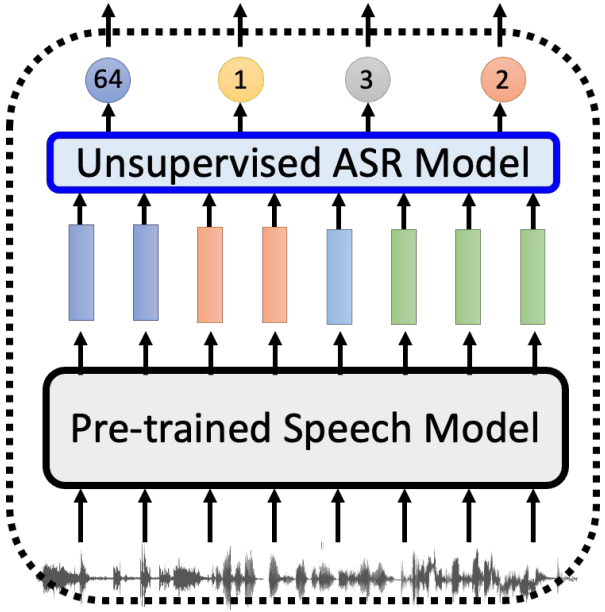
Toolkit

- Prosody-related Tasks
- Spoken Language Understanding

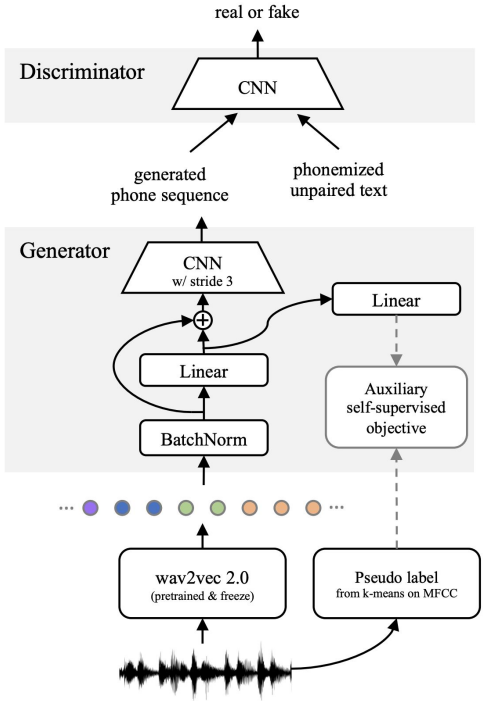
Adopt Unsupervised ASR as a Pre-trained Model

- Why unsupervised ASR can be a pre-trained model?
  - Unsupervised ASR does not use supervised data (i.e., no parallel data of speech and text)

- Potential benefits from unsupervised ASR
  - By training with unsupervised ASR objective, we expect to learn linguistical information from external text sources



Background - Unsupervised ASR



- Goal:
  - ASR for every languages without speech annotation
- End-to-end approach:
  - Speech SSL model
  - Adversarial objective

Towards End-to-end Unsupervised Speech Recognition (Liu et al. 2022)

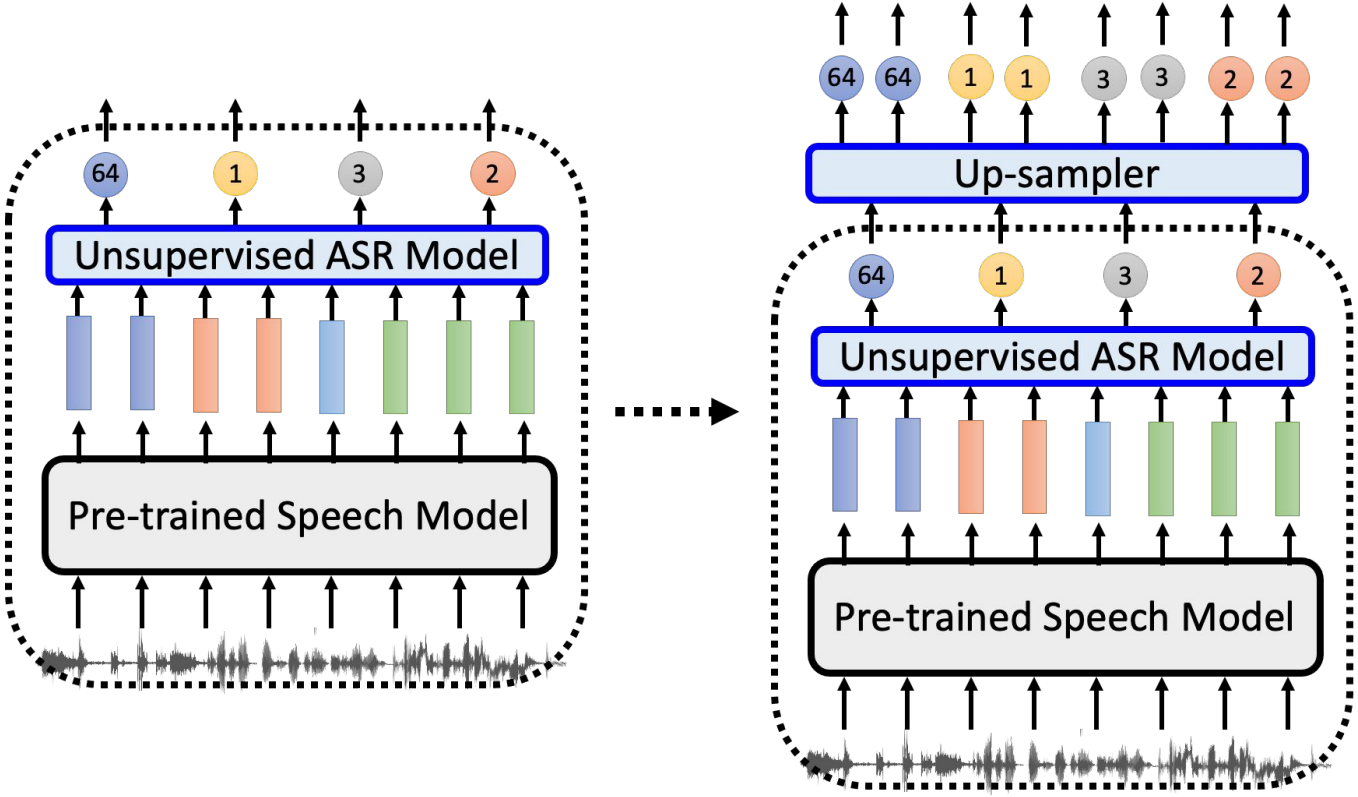
Utilize Unsupervised ASR Model to Various Directions

1. Adopt unsupervised ASR as a pre-trained model
   - Use unsupervised ASR directly for downstream task

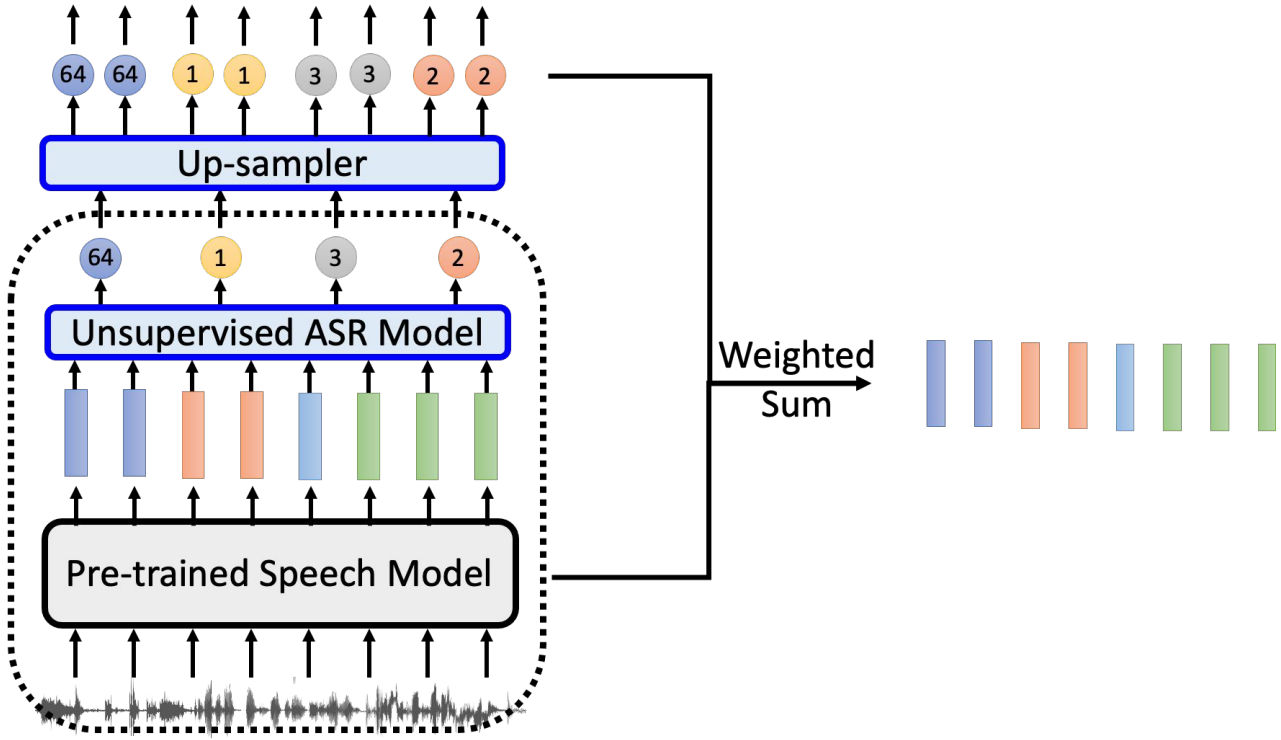
2. Connect unsupervised ASR to text-pretrained models
   - Get better connection between Speech SSL models to text SSL models

3. Utilize unsupervised ASR as a segmenter
   - Use unsupervised ASR to get phoneme-level segmentation in unsupervised fashion

Adopt Unsupervised ASR as a Pre-trained Model



Adopt Unsupervised ASR as a Pre-trained Model



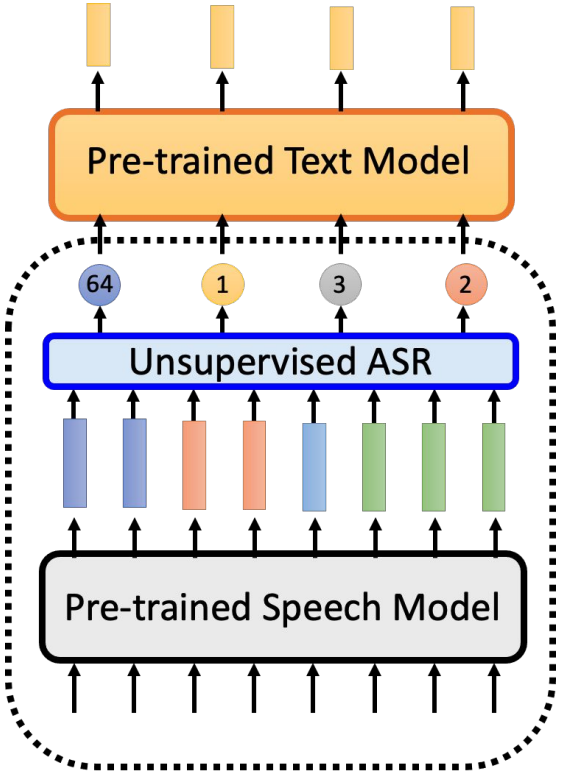
Adopt Unsupervised ASR as a Pre-trained Model

SSL Model	F1 in Superbbenchmark (SF)
Fbank (baseline)	64.18%
Wav2vec2	86.88%
Hubert	88.50%
Unsupervised ASR (with Wav2vec2)	87.75%

We are in progress of evaluating the performance on more tasks (e.g., ASR, IC, etc.) and more variants of unsupervised ASR (e.g., with different SSL front-ends)

Connect Unsupervised ASR to Textual Pre-trained Models

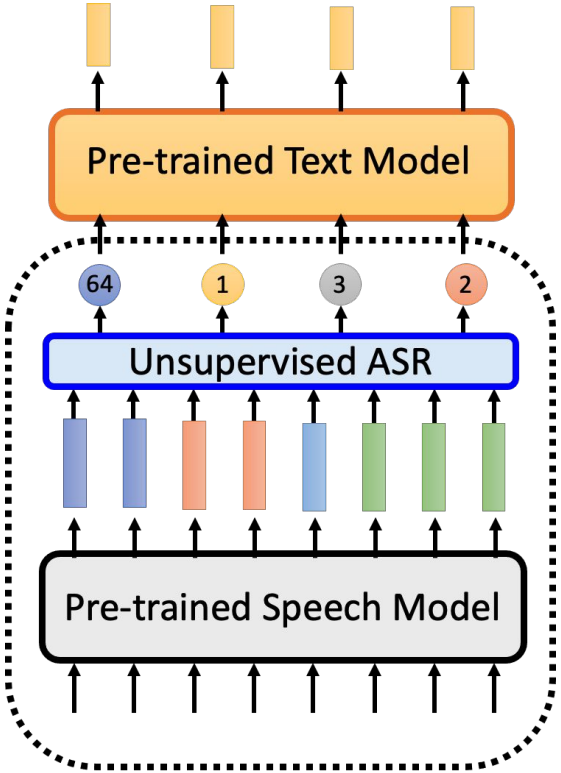
- Initial thoughts
  - We want to build BPE-based unsupervised ASR to convert unsupervised ASR for pre-trained text models
  - Failed due to difficulties in training unsupervised ASR for BPE



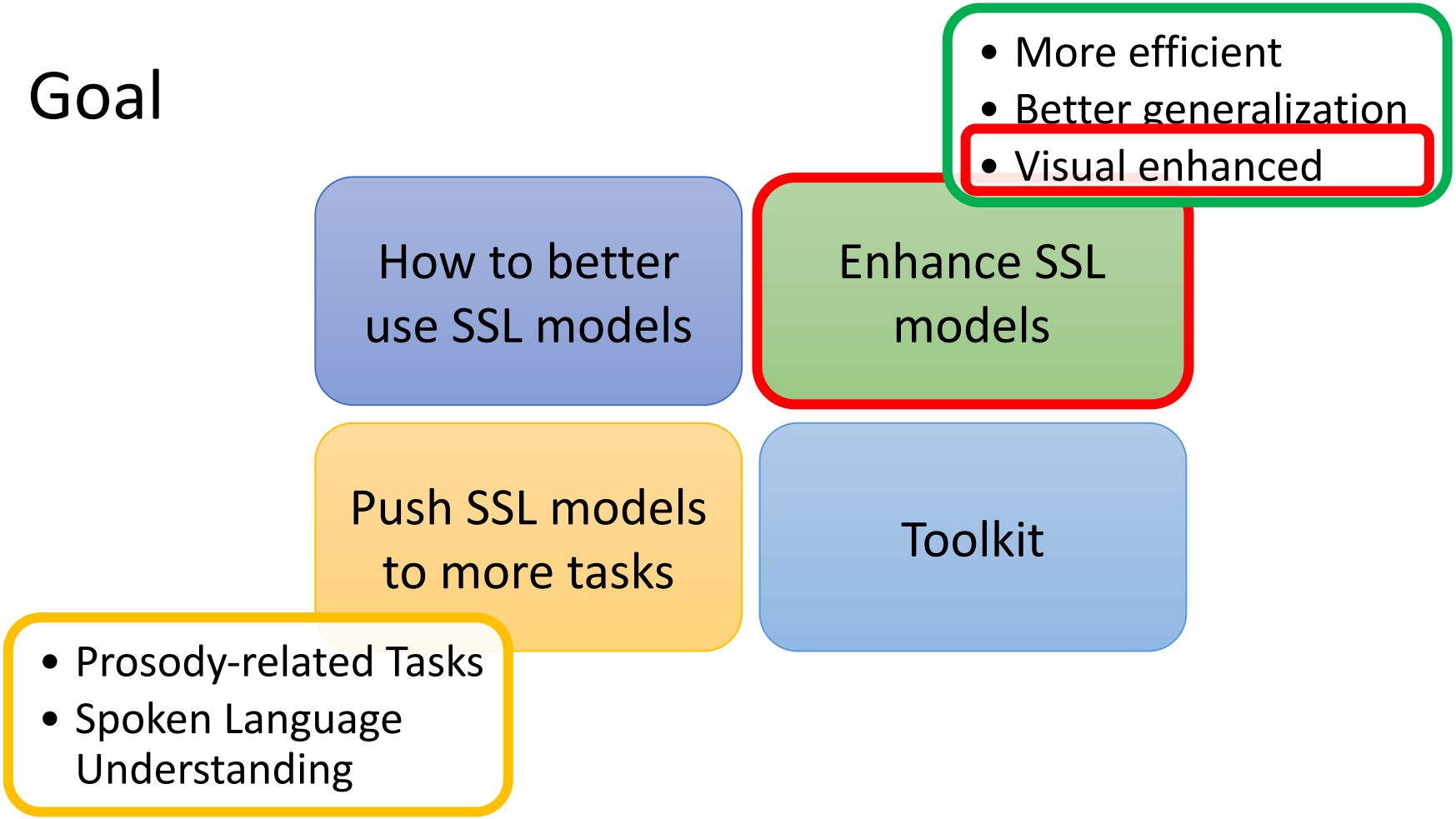
Connect Unsupervised ASR to Textual Pre-trained Models

Second thoughts:

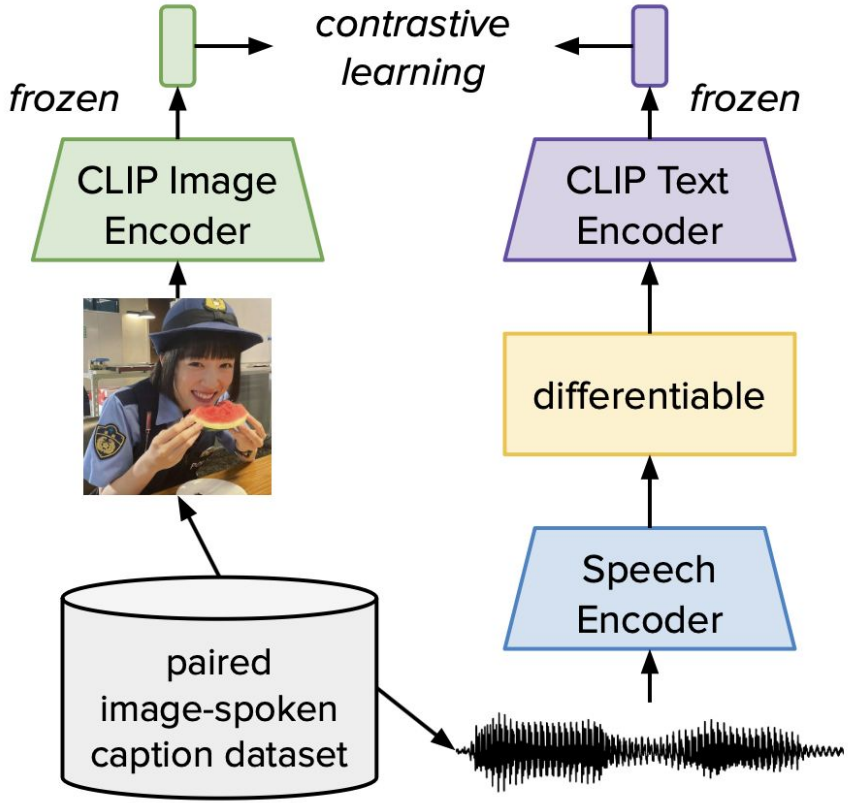
- Use phoneme-based pre-trained text model?
- We luckily have one there
  - https://huggingface.co/voidful/phoneme_byt5_v2
  - which could reach comparable performances to BPE-based model on GLUE
- The work of combination is still in progress



Goal


How to better use SSL models
Enhance SSL models
Push SSL models to more tasks
Toolkit
• More efficient
• Better generalization
• Visual enhanced
• Prosody-related Tasks
• Spoken Language Understanding

Visually-Enhanced Self-Supervised Speech Models



Speech Encoder Pre-training

- Goal:
  - Make speech encoder learn better **content & semantics** representations.
- Method:
  - Relate outputs of speech encoder and outputs of pretrained CLIP Text encoder.
- Downstream task:
  - Content/semantics related downstream tasks (ASR, IC, KS)
  - Sth to do with Subword Embedding

picture source: TA 張恆瑞

Visually-Enhanced SSL Model Team

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

Students:

NTU	**UT Austin**	**JHU**
Heng-jui Chang	Layne Berry	Elizabeth Boroda
Ian Shih		
Jeff Wang		

Motivation

- It is relatively easy to collect text image captions
- It is also relatively easy to collect spoken image captions (people just describe what they see, $0.03-0.04 per utterance or $0.18-$0.24 per minute)
- It is more expensive to collect transcribed speech data (collect speech first, then transcribe it at $1.50-$3.00/minute)
- Spoken image descriptions also provide a semantic training signal for languages that do not have a standard orthographic representation (Swiss German, Egyptian Arabic, many more…)
- Lots of recent progress on big pre-trained models:
  - Image-text matching models (e.g. CLIP)
  - (Unimodal) self-supervised speech models (e.g. wav2vec2.0, HuBERT)
  - (Multimodal) self-supervised speech-vision models (e.g. DAVEnet, FaST-VGS, VG-HuBERT)
- **Can we combine all of the above to get better performance on tasks like SUPERB or semantic speech-image retrieval?**

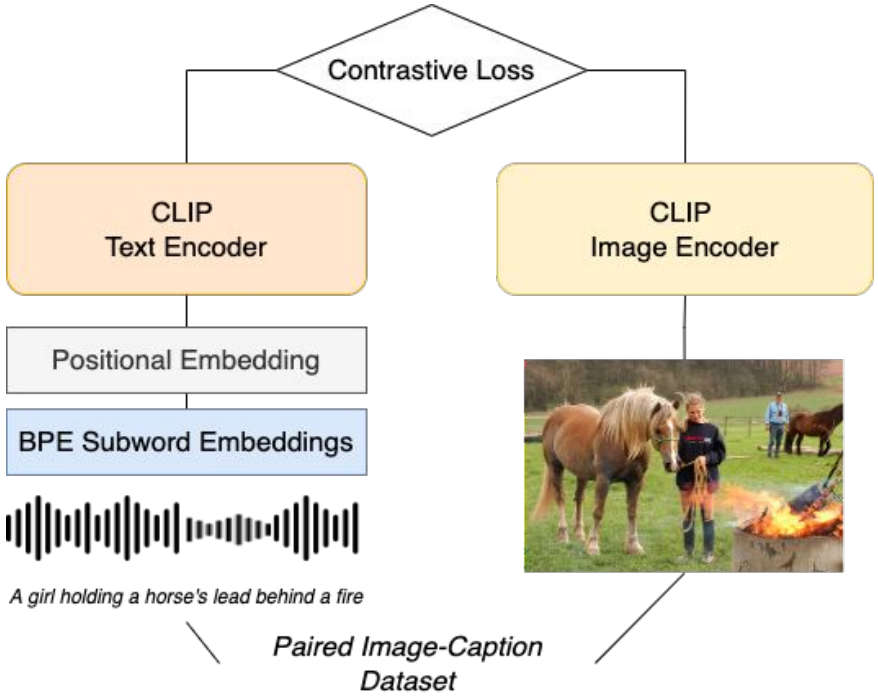
Image Caption Datasets

- Flickr 8k
  - Train: 6K Images 30K Captions (text/audio)
  - Dev: 1K Images 5K Captions (text/audio)
  - Test: 1K Images 5K Captions (text/audio)
- SpokenCOCO
  - Train: 113K Images 565k Captions (text/audio)
  - Dev: 5k images 25k Captions (text/audio)
  - Test: 5k images 25k Captions (text/audio)

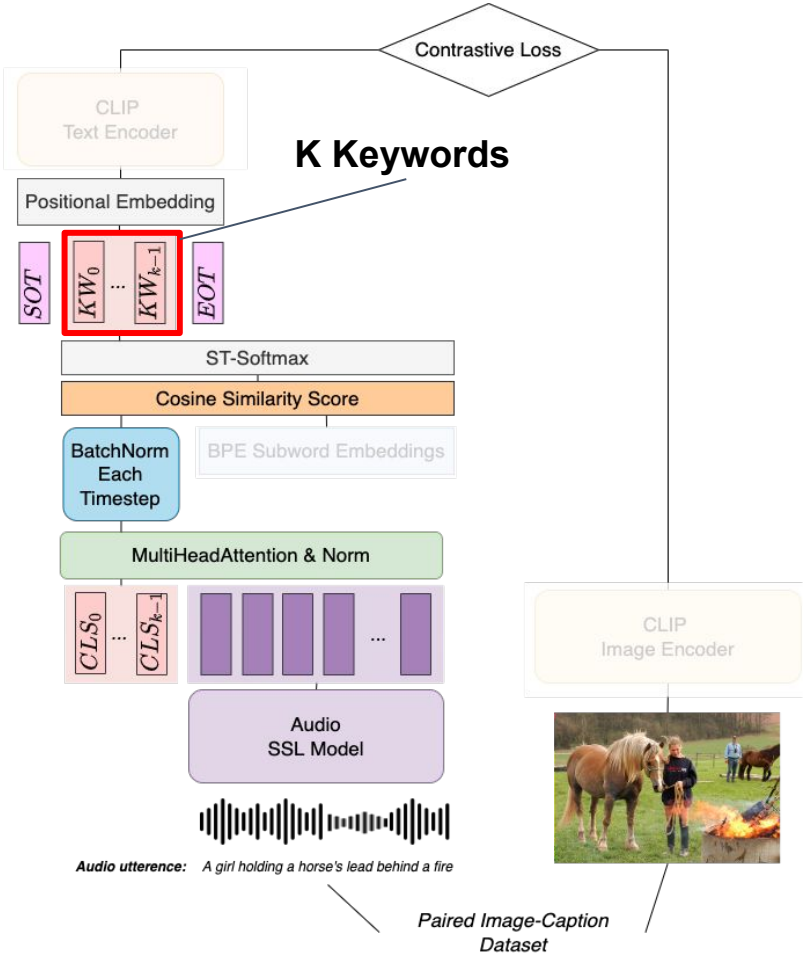
Example:

A blonde horse and a blonde girl in a black sweatshirt are staring at a fire in a barrel .
A girl and her horse stand by a fire .
A girl holding a horse 's lead behind a fire .
A man , and girl and two horses are near a contained fire .
Two people and two horses watching a fire .

Model Structure

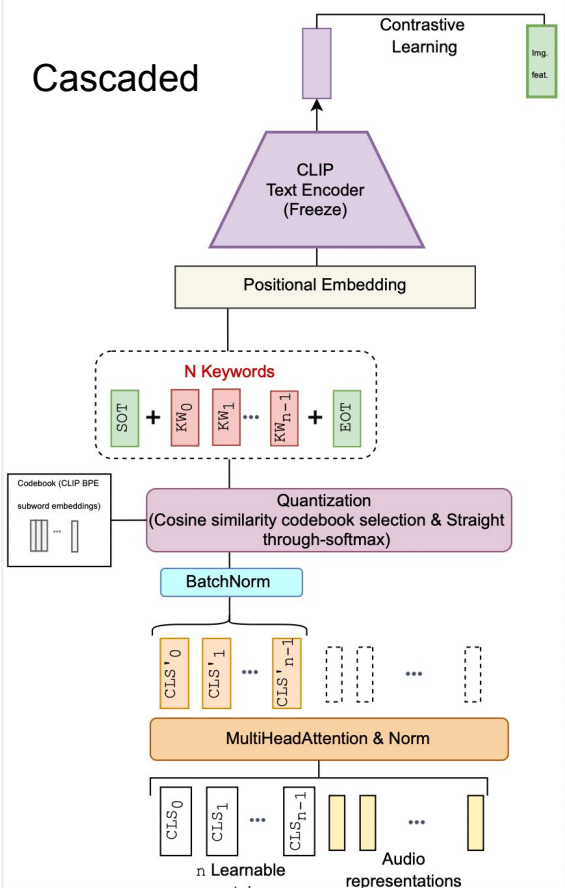


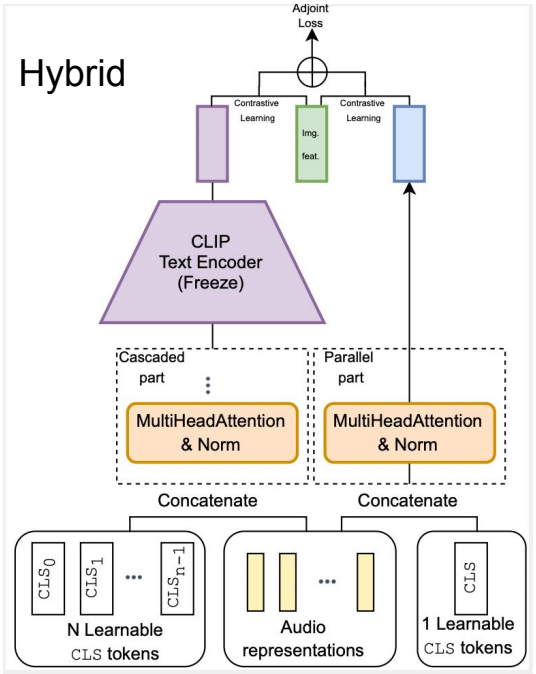
Original CLIP Setting

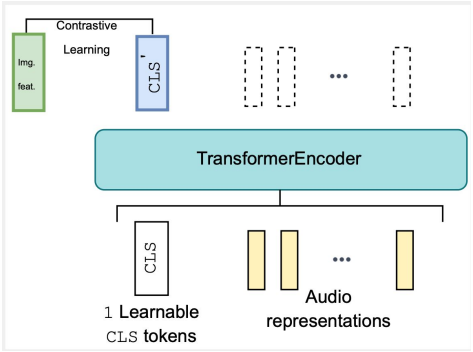


Cascaded CLIP Setting

Three Model Variants







Evaluation

1. **SUPERB**
   a. Primarily looking at ASR and PR, but also KS, QbE, IC, SF

1. **Speech-Image retrieval**
   a. Goal: given a library of N images and a spoken description of one of those images as input, search over the library and return the target image being described
   b. Results reported with Recall @ top 1, 5, 10 returned results
   c. Can also search in the opposite direction (images -> speech)

Speech-Image Retrieval Results

Table 1. Table to test captions and labels. **WS** indicates weight sharing, meaning that the parallel and cascaded parts of the integrated SpeechCLIP shared the same attention layer.

Model	Speech → Image			Image → Speech		
	R@1	R@5	R@10	R@1	R@5	R@10
	Flickr8k (Test)					
Cascaded	8.2	25.7	37.2	14.1	34.5	49.2
Parallel	26.7	57.1	70.0	41.3	73.9	84.2
Hybrid_{C}	6.1	19.7	30.0	8.0	25.7	39.7
Hybrid_{P}	24.9	55.7	69.0	37.3	67.8	81.4
$\text{Hybrid-WS}_{\text{C}}$	6.8	20.9	32.4	10.3	29.0	41.8
$\text{Hybrid-WS}_{\text{P}}$	27.0	56.8	69.5	41.1	73.0	83.5
Cascaded Large	6.4	20.7	31.0	9.6	27.7	39.7
Parallel Large	35.8	66.5	78.0	50.6	80.9	89.1
Hybrid_{C} Large	7.3	21.3	31.8	10.9	29.1	41.0
Hybrid_{P} Large	37.0	67.1	78.3	51.8	80.5	89.3
$\text{Hybrid-WS}_{\text{C}}$ Large	4.0	14.1	22.4	6.5	19.8	30.1
$\text{Hybrid-WS}_{\text{P}}$ Large	33.5	61.9	73.9	44.9	75.0	84.9
$\text{FaST-VGS}_{\text{CO}}$	26.6	56.4	68.8	36.2	66.1	76.5
$\text{FaST-VGS}_{\text{CTF}}$	29.3	58.6	71.0	37.9	68.5	79.9
CLIP	1	5	10	1	5	10

Model	Speech → Image			Image → Speech		
	R@1	R@5	R@10	R@1	R@5	R@10
	SpokenCOCO (Test)					
Cascaded Large	14.7	41.2	55.1	21.8	52.0	67.7
Parallel Large	39.1	72.0	83.0	54.5	84.5	93.2
Hybrid_{C} Large	15.2	39.8	54.6	19.2	50.2	63.5
Hybrid_{P} Large	36.6	66.9	77.8	48.3	82.5	90.1
$\text{Hybrid-WS}_{\text{C}}$ Large	7.6	24.7	36.6	10.6	30.8	45.3
$\text{Hybrid-WS}_{\text{P}}$ Large	42.7	74.8	84.4	55.3	85.7	94.0
$\text{FaST-VGS}_{\text{CO}}$	31.8	62.5	75.0	42.5	73.7	84.9
$\text{FaST-VGS}_{\text{CTF}}$	35.9	66.3	77.9	48.8	78.2	87.0
CLIP	1	5	10	1	5	10

SUPERB Results

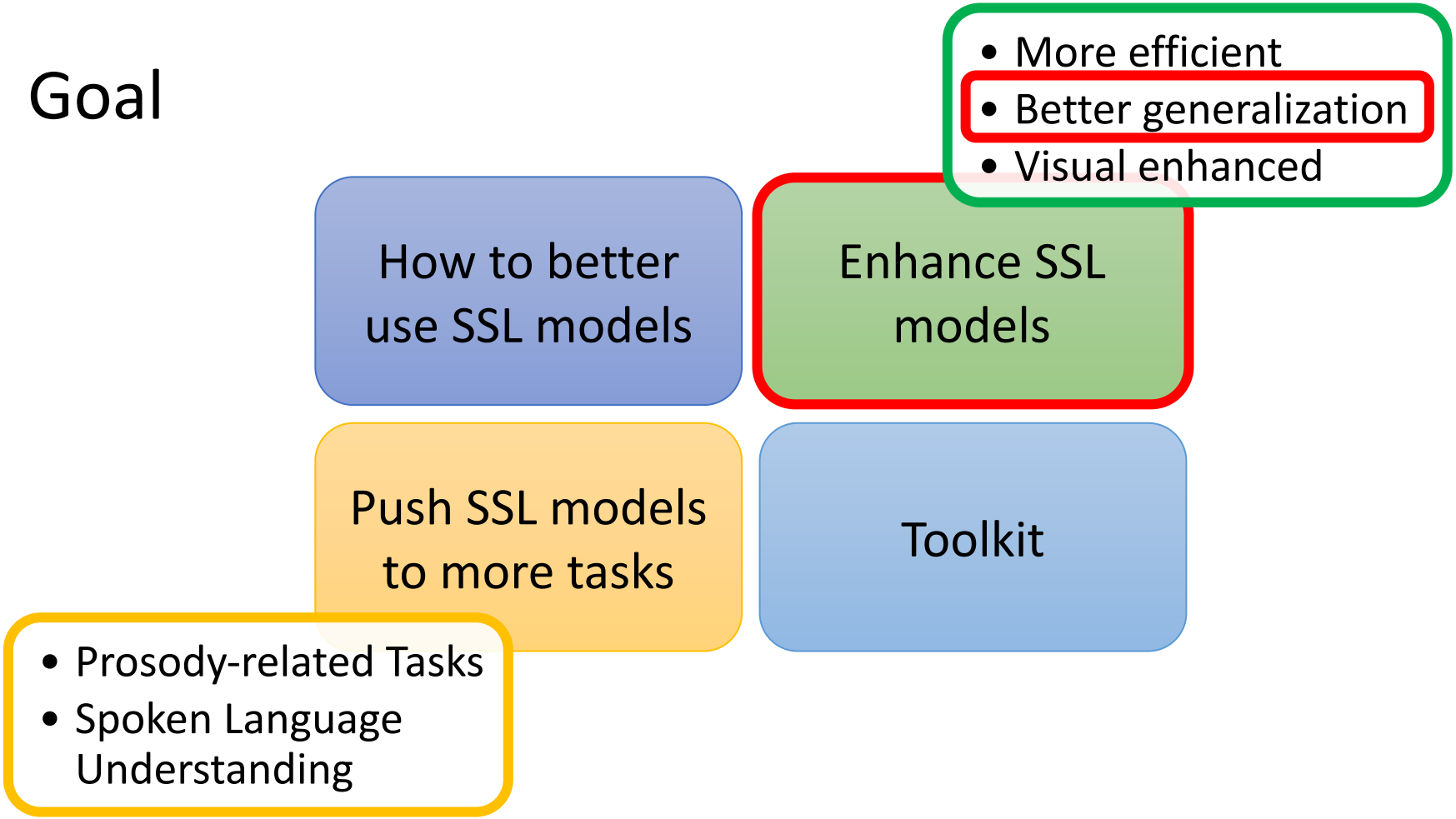
			Content				Semantics	
			PR	ASR (+LM)	KS	QbE	IC	SF
Method	# Param.	Data	PER↓	WER↓	Acc↑	MTWV↑	Acc↑	F1↑ / CER↓
Speech SSL Baselines								
HuBERT Base	94.68M	LS960	5.41	6.42 / 4.79	96.30	0.0736	98.34	88.53 / 25.20
HuBERT Large	316.61M	LL60k	3.53	3.62 / 2.94	95.29	0.0353	98.76	89.81 / 21.76
WavLM Large	316.62M	Mix94k	3.06	3.44 /	97.86	0.0886	99.31	92.21 / 18.86
Visually Enhanced Speech SSL								
FaST-VGS	187.87M	LS960+SC742	16.30	13.46 / 9.51	96.85	0.0546	98.37	84.91 / 32.33
FaST-VGS+	217.23M	LS960+SC742	7.76	8.83 / 6.37	97.27	0.0562	98.97	88.15 / 27.12
SpeechCLIP (Ours)								
Cascaded		LS960+YFCC15M+FACC	4.92	6.34	96.62		98.02	88.90 / 23.28
Parallel		LS960+YFCC15M+FACC	4.95	6.46	96.56		97.71	88.23 / 24.93
Hybrid		LS960+YFCC15M+FACC	5.09	6.36	96.66		98.37	88.73 / 24.58
Hybrid-WS		LS960+YFCC15M+FACC	4.98	6.47	96.59		98.15	88.26/ 25.04
Cascaded Large		LL60k+YFCC15M+FACC			95.62		98.52	
Parallel Large		LL60k+YFCC15M+FACC	2.94	3.68	95.68		98.50	89.18 / 22.38
Hybrid Large		LL60k+YFCC15M+FACC			95.62		98.60	
Hybrid-WS Large		LL60k+YFCC15M+FACC			95.55		98.71	89.42 / 21.99
Cascaded Large		LL60k+YFCC15M+SC742	2.97		95.62		98.52	
Parallel Large		LL60k+YFCC15M+SC742	2.89	3.66	95.62		98.73	
Hybrid Large		LL60k+YFCC15M+SC742	3.00		95.42			88.98 / 22.77
Hybrid-WS Large		LL60k+YFCC15M+SC742			95.59			88.91 / 22.80

Plan for ongoing work

- Submit SpeechCLIP results to SLT 2022 (July 21)

- Parallel direction: improving unsupervised ASR with visual information
  - Integrating VG-HuBERT and CLIP into wav2vec-U
  - Will give an update on this direction next time

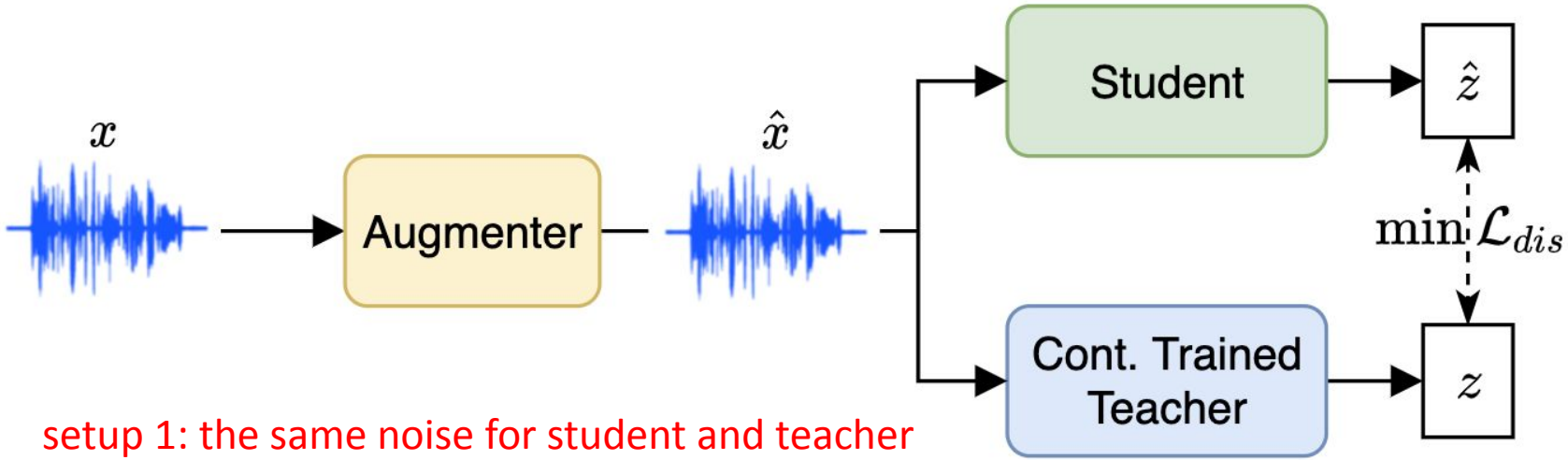
Goal


How to better use SSL models
Enhance SSL models
Push SSL models to more tasks
Toolkit
• More efficient
• Better generalization
• Visual enhanced
• Prosody-related Tasks
• Spoken Language Understanding

Continually trained teacher with distorted input

***Stage 1*: Continually train** the teacher model with distorted data.

***Stage 2*:** Knowledge distillation



setup 1: the same noise for student and teacher

setup 2: different noises for student and teacher

Results

	Intent Classification			Emotion Recognition		Keyword Spotting	
Testing Data	clean	m+g+r	fsd	clean	m+g+r	clean	m+g+r
DistilHuBERT	94.78	66.41	-	63.87	53.92	96.04	89.84
+ setup 1	95.39	81.97	86.98	64.98	57.05	96.43	95.00
+ setup 2	96.18	88.48	91.25				

	Speaker Identification		ASR (WER)	
Testing Data	clean	m+g+r	clean	m+g+r
DistilHuBERT	73.02	40.42	13.77	37.59
+ setup 1	73.13	57.34	13.25	19.17