NAACL 2022

Cross-modal Contrastive Learning for Speech Translation

Rong Ye, Mingxuan Wang, Lei Li

Paper:

Code:

字节跳动 ByteDance

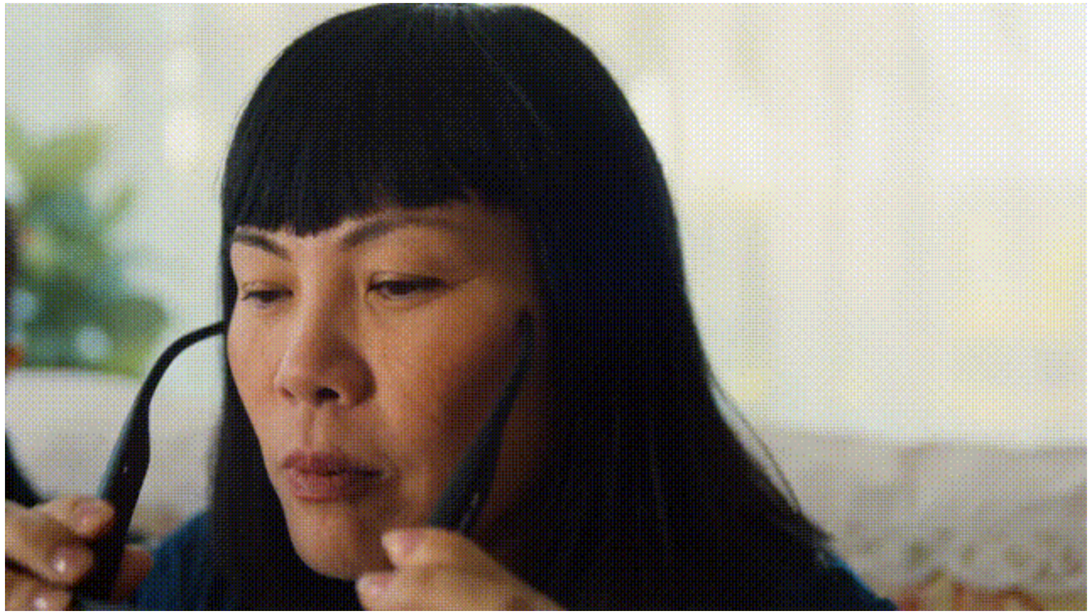
UC SANTA BARBARA

Speech-to-text **T**ranslation (ST)

Source language ***speech(audio)***
→ Target language ***text***

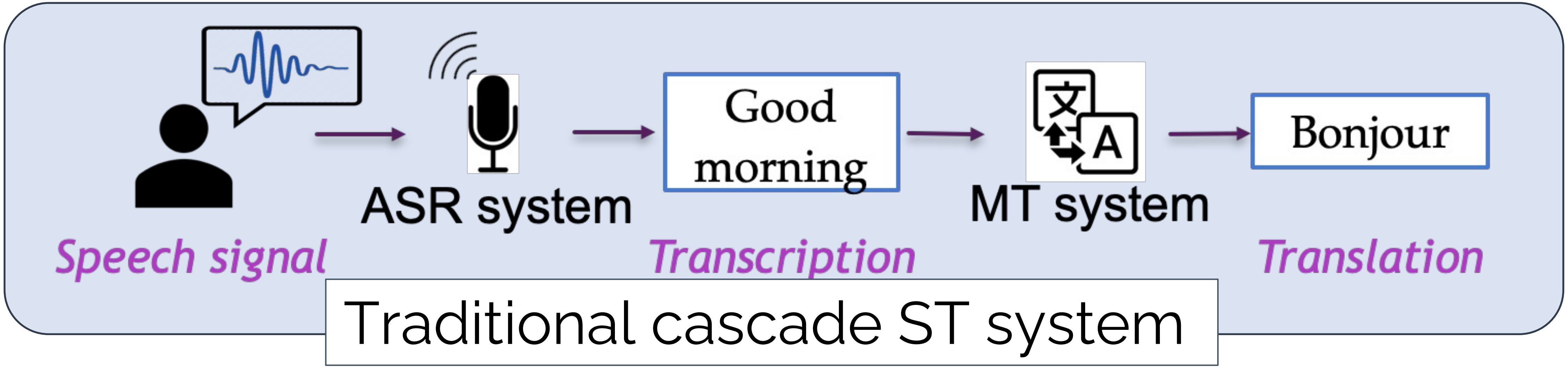
Wide Applications of ST

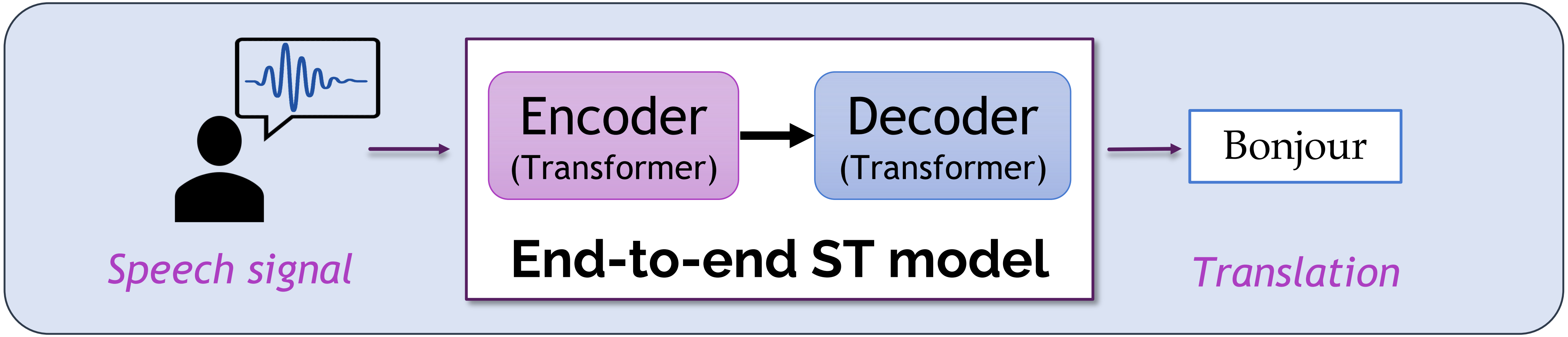
English subtitle for popular Korean TV drama *"Squid Game"*



Google Live AR Translation Service Glasses Prototype Google I/O 2022

End-to-end model: makes ST easier



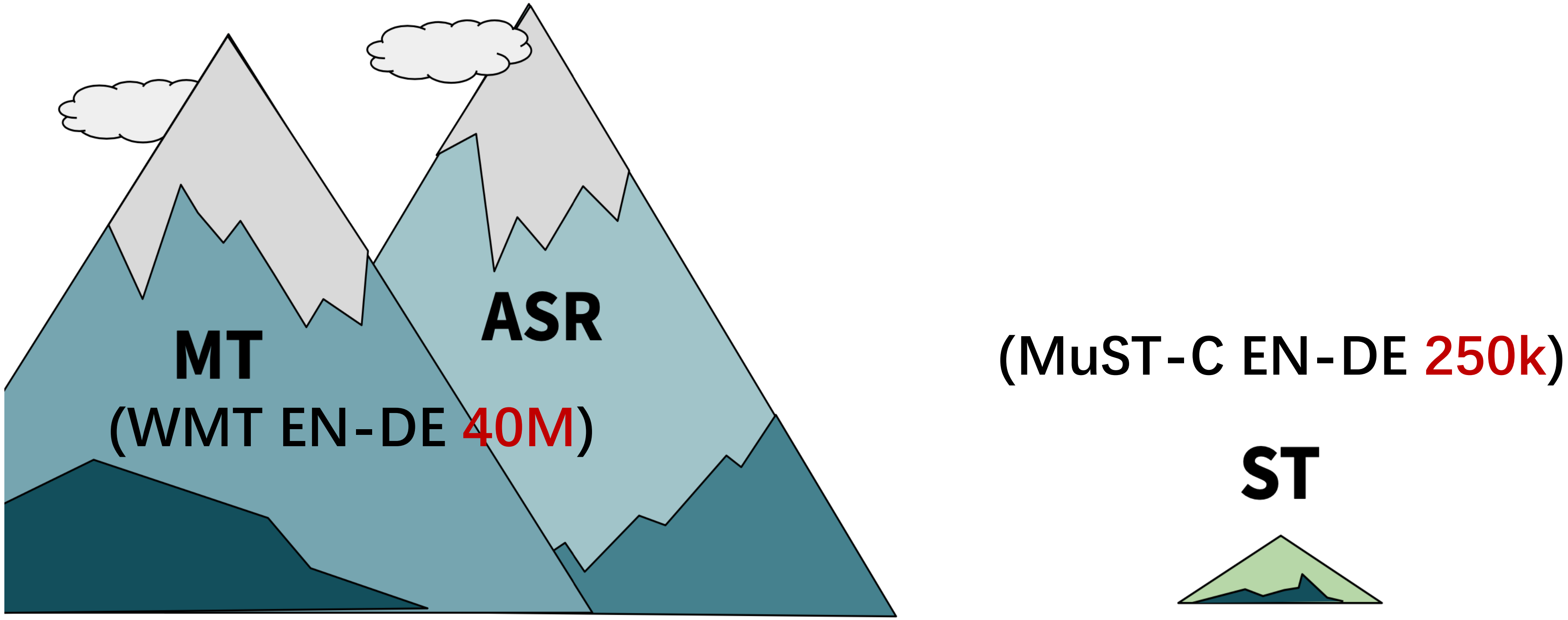


* Pictures are from our previous video talk at InterSpeech 2021.

Why end-to-end ST is hard?

- **Data Scarcity** - lack of large parallel corpus

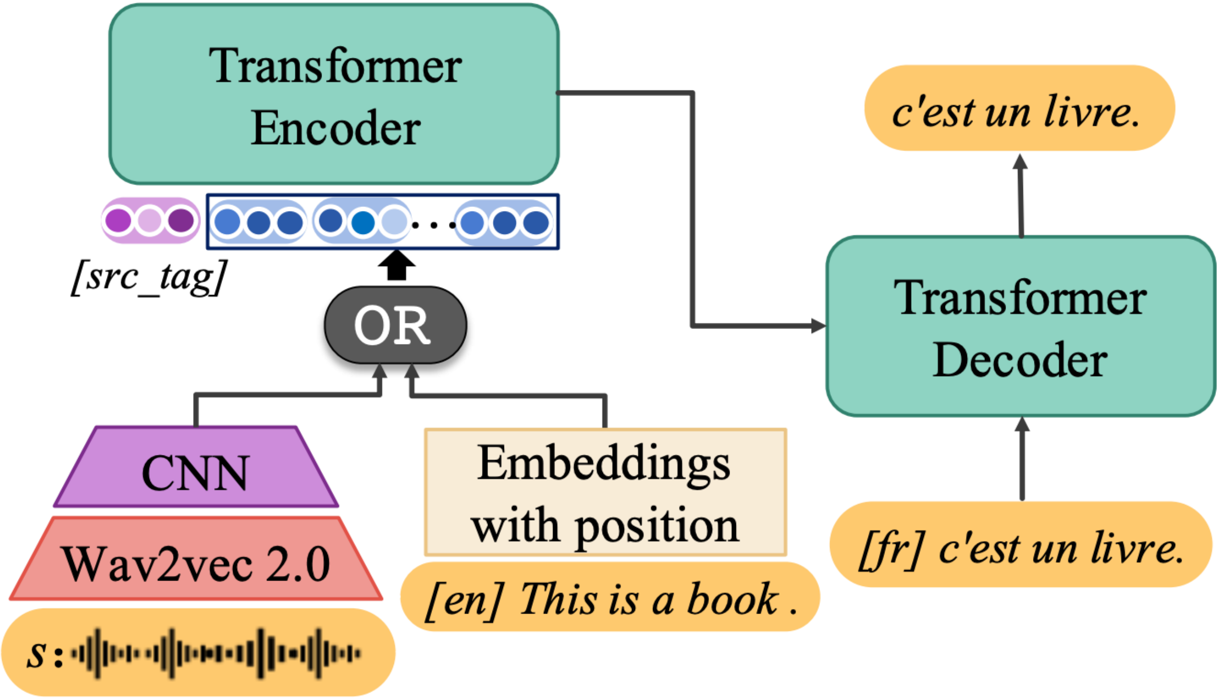
<speech, transcript_text, translate_text>



* Pictures are from previous EACL 2021 tutorial on speech translation. https://st-tutorial.github.io/

Multi-task learning leads to better ST

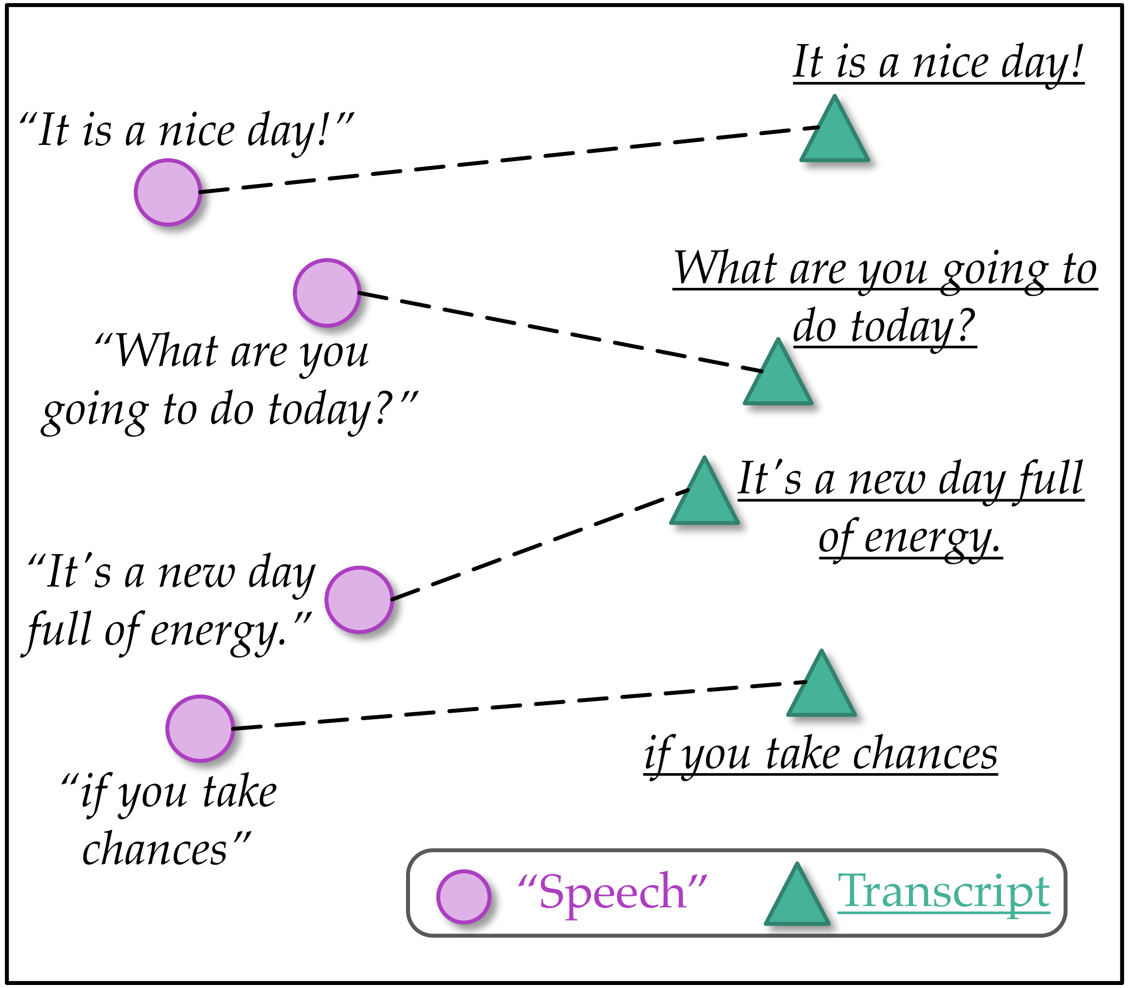
- To joint train ST, ASR and MT tasks.
- **Advantages**:
  - Better generalization
  - Utilizing large-scale extra MT, ASR data.

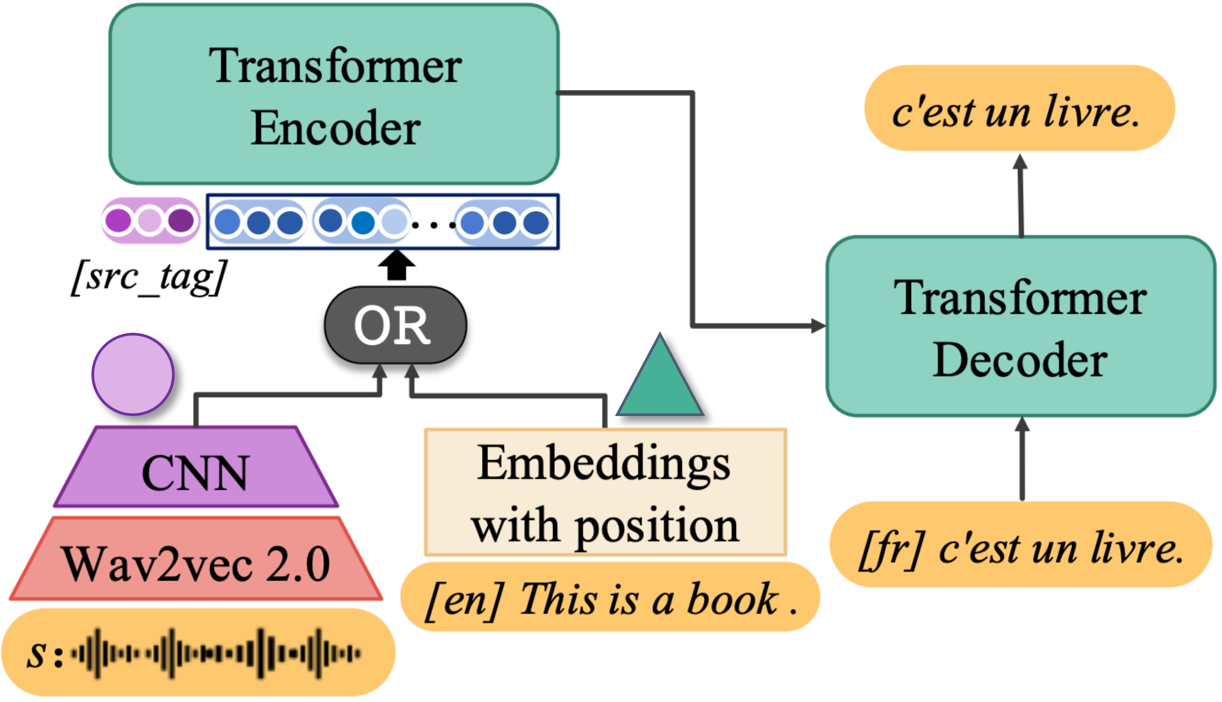


XSTNet (Ye et al., 2021[1])

[1] Rong Ye, Mingxuan Wang, and Lei Li. XSTNet: End-to-end Speech Translation via Cross-modal Progressive Training. InterSpeech 2021.

Representation Perspective: Modality **Gap** Exists!

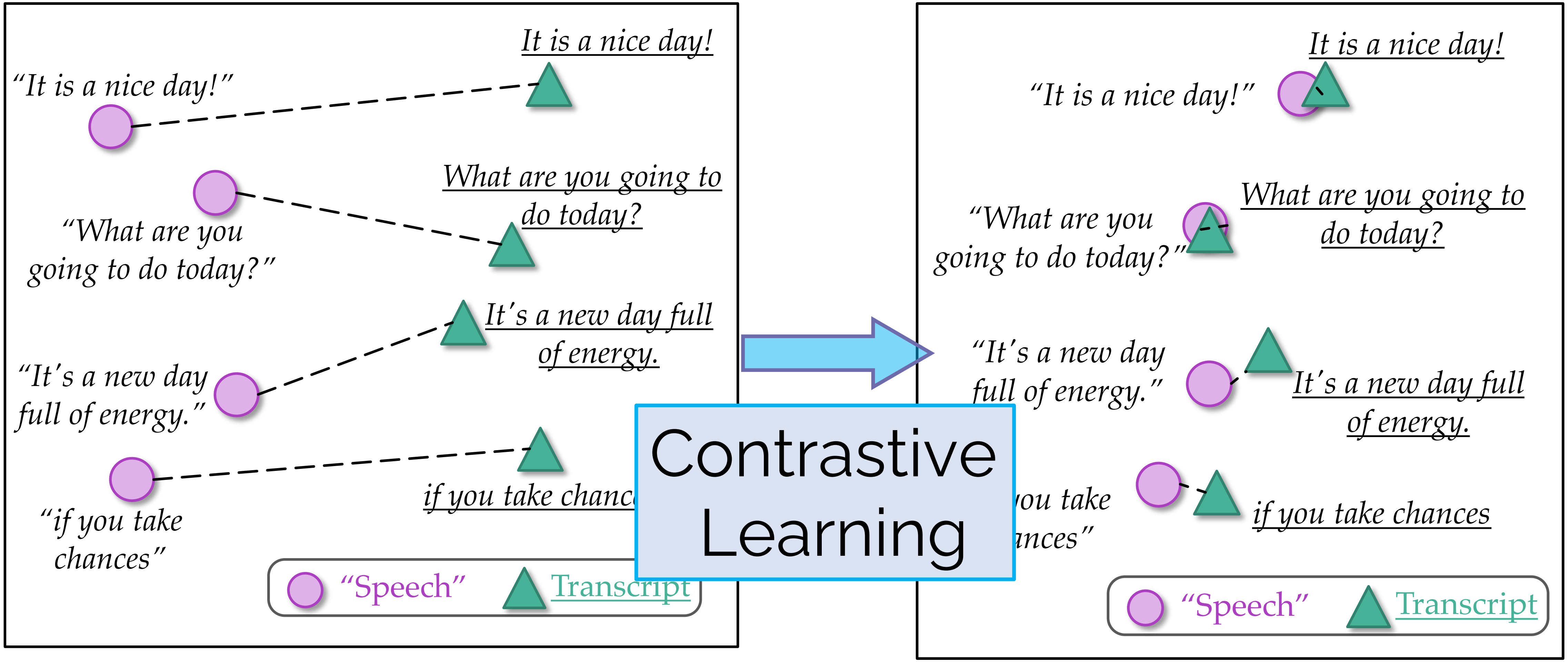




XSTNet (Ye et al., 2021[1])

[1] Rong Ye, Mingxuan Wang, and Lei Li. XSTNet: End-to-end Speech Translation via Cross-modal Progressive Training. InterSpeech 2021.

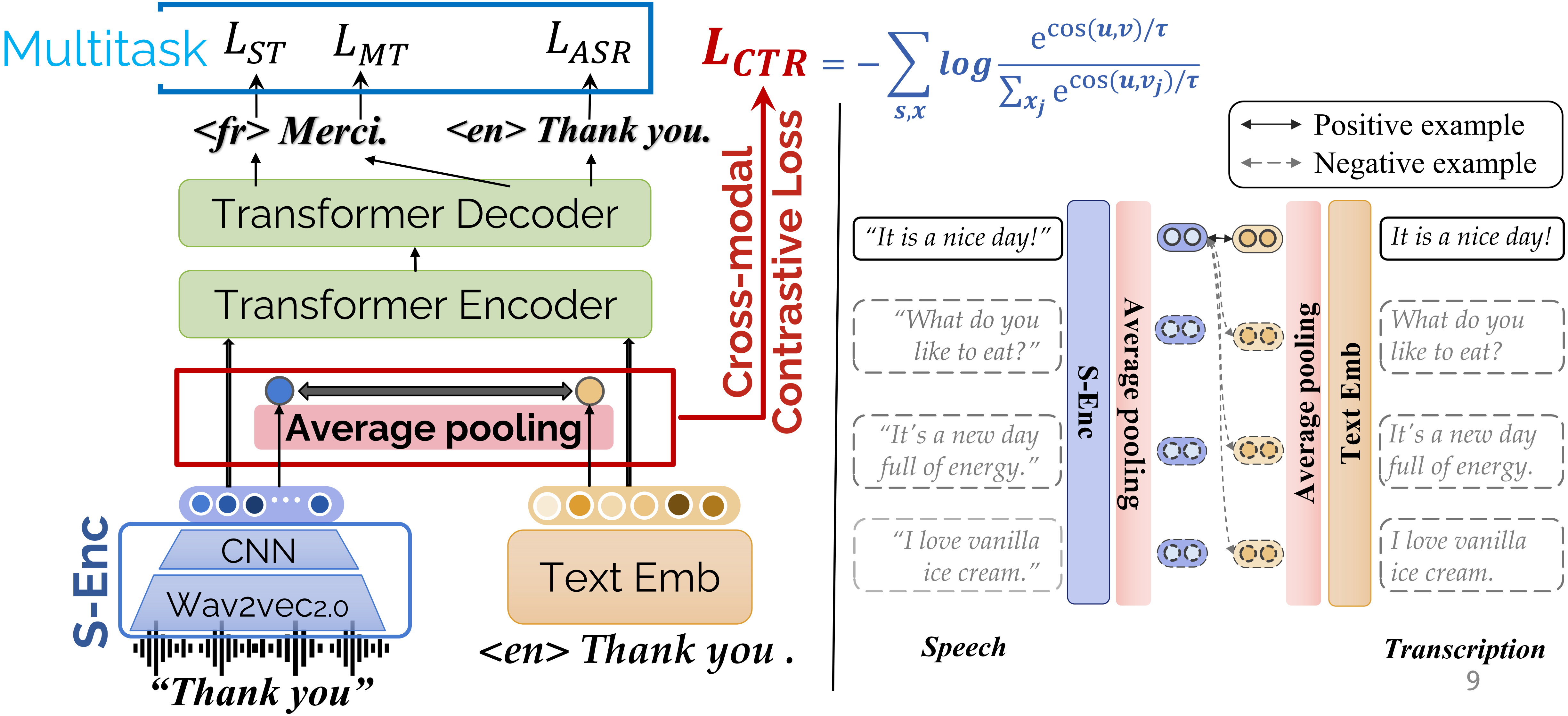
Text and speech with same meaning should be **similar** in representation!



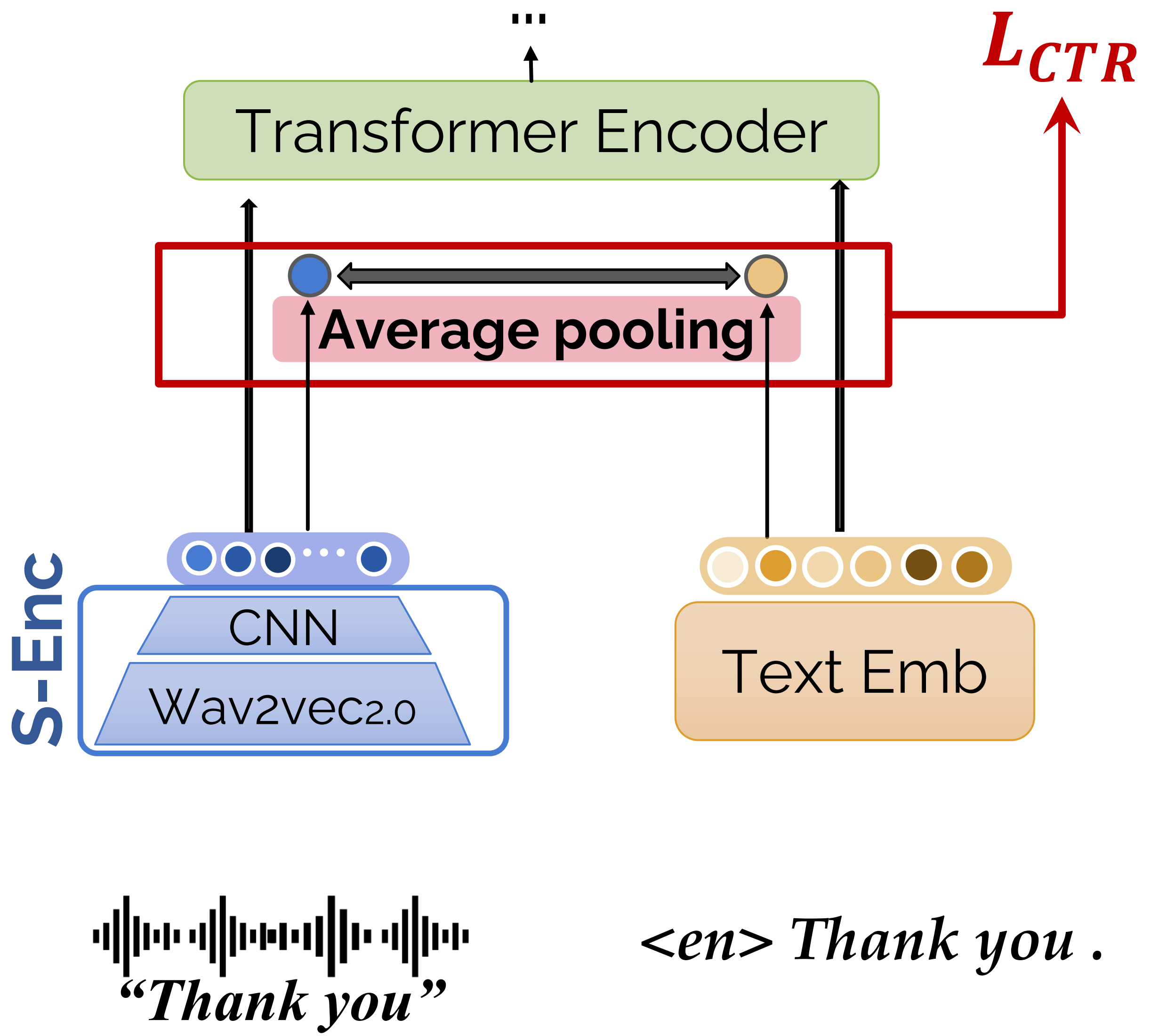
(a) Current models

(b) Expected

Method: **Con**trastive Learning (**Con**ST)



[Optional] Mining more hard examples



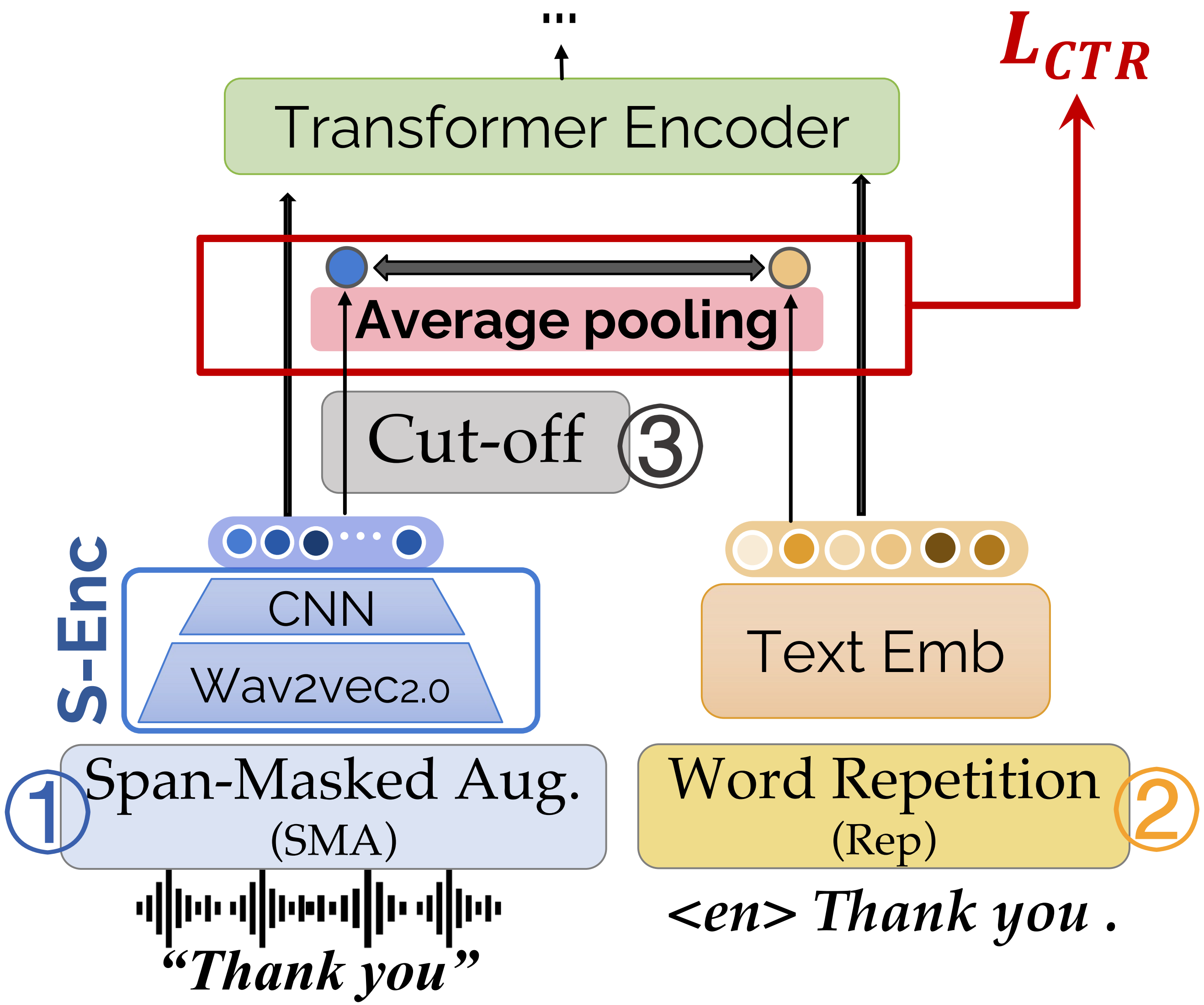
We introduce three hard example mining operations.

① Span-Masked Aug. (SMA)

② Word Repetition (Rep)

③ Cut-off

[Optional] Mining more hard examples



We introduce three hard example mining operations.

- Input level
- Representation level

Experiments

Experimental Setups

- **Datasets**
  - All 8 directions of **MuST-C** benchmark
  - MT datasets for pretraining

En→	ST (MuST-C) hours	ST (MuST-C) #sents	MT name	MT #sents
De	408	234K	WMT16	4.6M
Fr	492	280K	WMT14	40.8M
Ru	489	270K	WMT16	2.5M
Es	504	270K	WMT13	15.2M
Ro	432	240K	WMT16	0.6M
It	465	258K	OPUS100	1.0M
Pt	385	211K	OPUS100	1.0M
Nl	442	253K	OPUS100	1.0M

- **Settings**
  - without external MT data
  - with external MT data

- **Baseline**
  - W2v2-Transformer
  - XSTNet (Ye et. al.)[1]

[1] Rong Ye, Mingxuan Wang, and Lei Li. XSTNet: End-to-end Speech Translation via Cross-modal Progressive Training. InterSpeech 2021.

Contrastive Learning improves ST

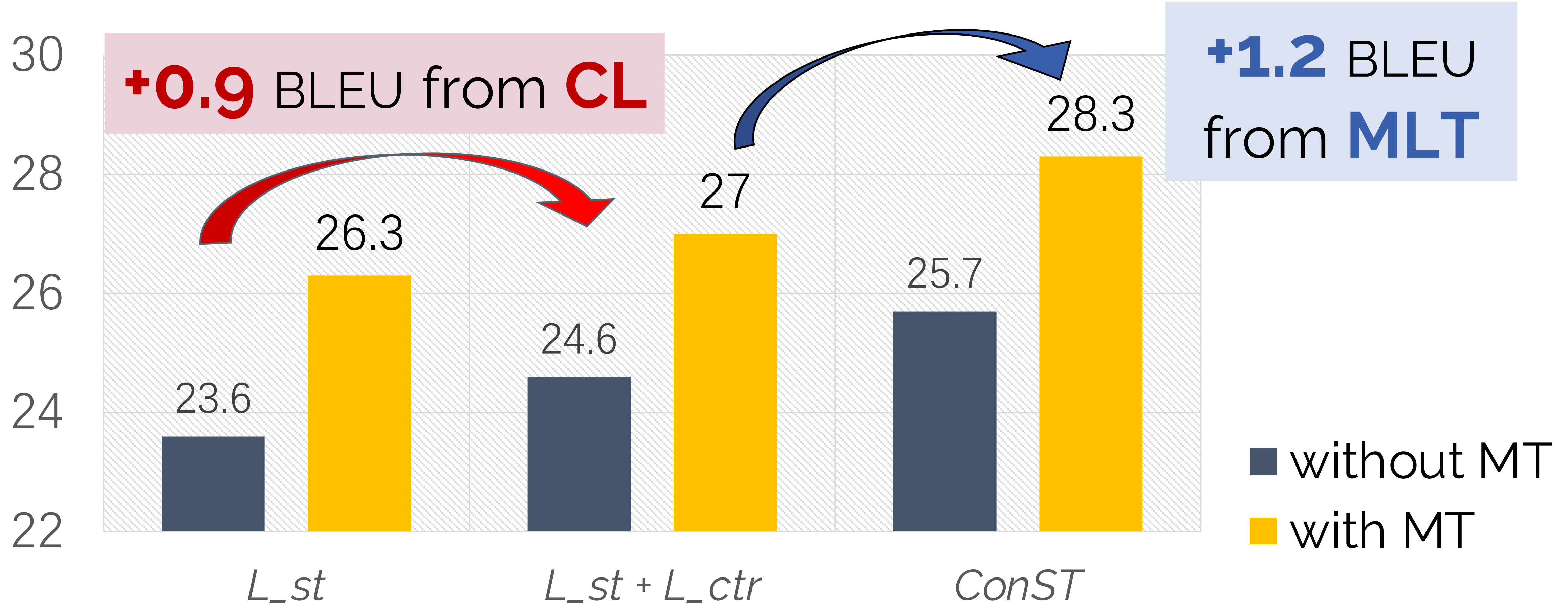
Models	External Data				BLEU								
	Speech	Text	ASR	MT	De	Es	Fr	It	Nl	Pt	Ro	Ru	Avg.
w/o external MT data													
Fairseq ST (Wang et al., 2020a)	-	-	-	-	22.7	27.2	32.9	22.7	27.3	28.1	21.9	15.3	24.8
NeurST (Zhao et al., 2021a)	-	-	-	-	22.8	27.4	33.3	22.9	27.2	28.7	22.2	15.1	24.9
Espnet ST (Inaguma et al., 2020)	-	-	-	-	22.9	28.0	32.8	23.8	27.4	28.0	21.9	15.6	25.1
Dual Decoder (Le et al., 2020)	-	-	-	-	23.6	28.1	33.5	24.2	27.6	30.0	22.9	15.2	25.6
W-Transf. (Ye et al., 2021)	✓	-	-	-	23.6	28.4	34.6	24.0	29.0	29.6	22.4	14.4	25.7
Speechformer (Papi et al., 2021)	-	-	-	-	23.6	28.5	-	-	27.7	-	-	-	-
LightweightAdaptor (Le et al., 2021)	-	-	-	-	24.7	28.7	35.0	25.0	28.8	31.1	23.8	16.4	26.6
Self-training (Pino et al., 2020)	✓	-	✓	-	25.2	-	34.5	-	-	-	-	-	-
SATE (Xu et al., 2021)	-	-	-	-	25.2	-	-	-	-	-	-	-	-
BiKD (Inaguma et al., 2021)	-	-	-	-	25.3	-	35.3	-	-	-	-	-	-
Mutual-learning (Zhao et al., 2021b)	-	-	-	-	-	28.7	36.3	-	-	-	-	-	-
XSTNet (Ye et al., 2021)	✓	-	-	-	25.5	29.6	36.0	25.5	30.0	31.3	**25.1**	16.9	27.5
ConST	✓	-	-	-	**25.7**	**30.4**	**36.8**	**26.3**	**30.6**	**32.0**	24.8	**17.3**	**28.0**
w/ external MT data													
Chimera (Han et al., 2021)	✓	-	-	✓	27.1†	30.6	35.6	25.0	29.2	30.2	24.0	17.4	27.4
XSTNet (Ye et al., 2021)	✓	-	-	✓	27.1	30.8	38.0	26.4	31.2	32.4	**25.7**	18.5	28.8
STEMM (Fang et al., 2022)	✓	-	-	✓	28.7	31.0	37.4	25.8	30.5	31.7	24.5	17.8	28.4
ConST	✓	-	-	✓	**28.3**	**32.0**	**38.3**	**27.2**	**31.7**	**33.1**	25.6	**18.9**	**29.4**

+ 0.5 BLEU

+ 0.6 BLEU

Both **Multi-task** and **Contrastive** Learning are important!

$$\mathcal{L} = \boxed{\mathcal{L}_{\mathrm{ST}} + \mathcal{L}_{\mathrm{ASR}} + \mathcal{L}_{\mathrm{MT}}} + \lambda\mathcal{L}_{\mathrm{CTR}}$$



Three More things on CL

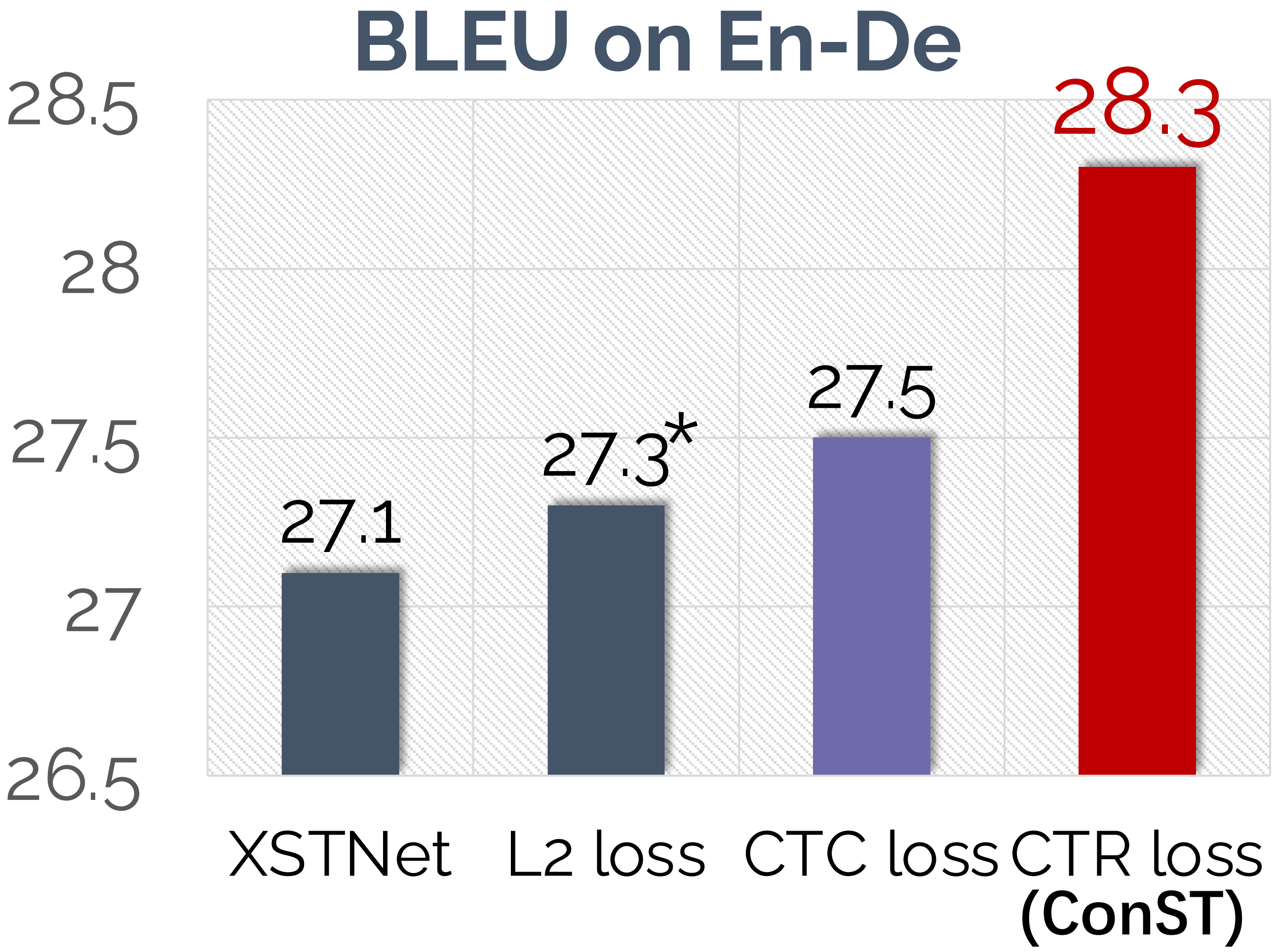
1. How effective are the hard examples mining operations? <u>(see paper, not in this slide)</u>

2. Is contrastive loss better than other losses?

3. Which layer to contrast on?

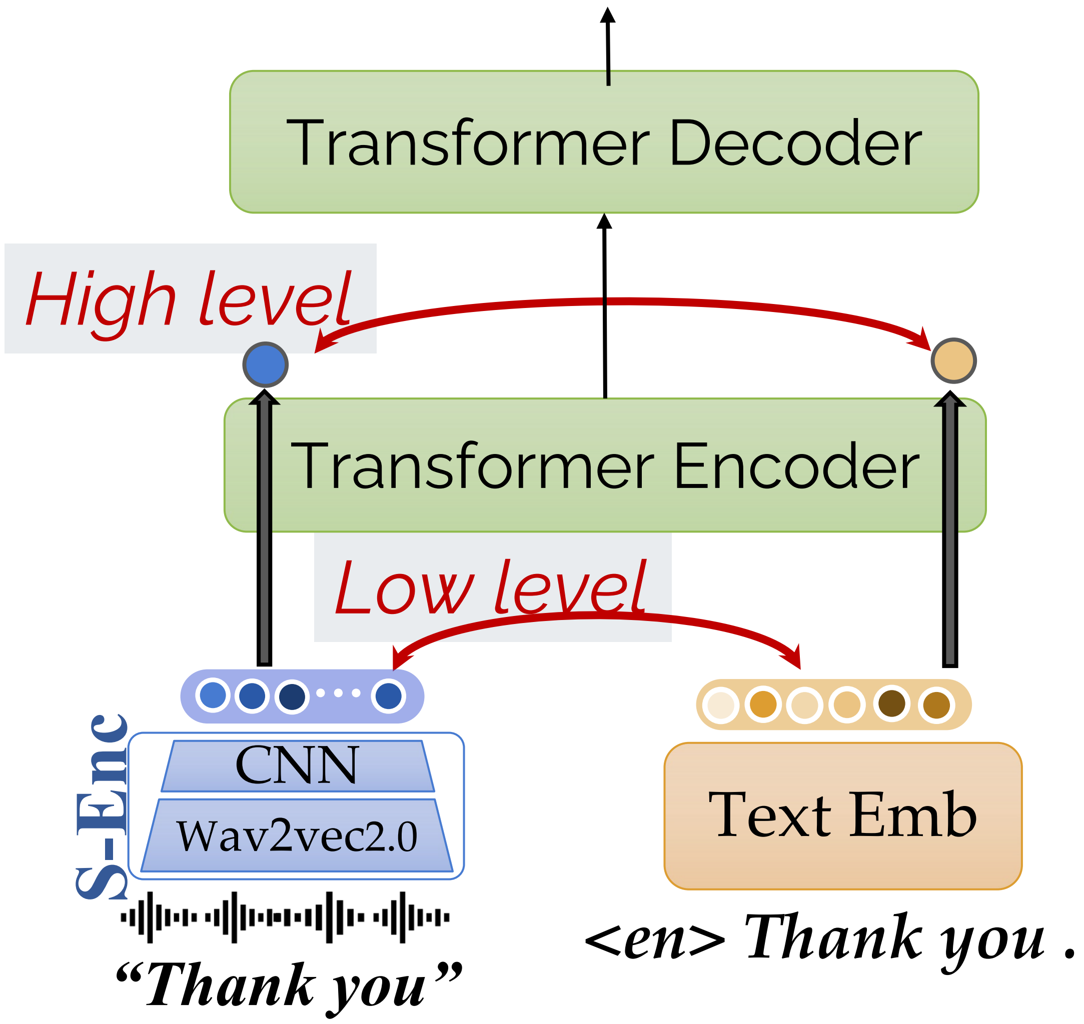
Contrastive loss: better than other losses!

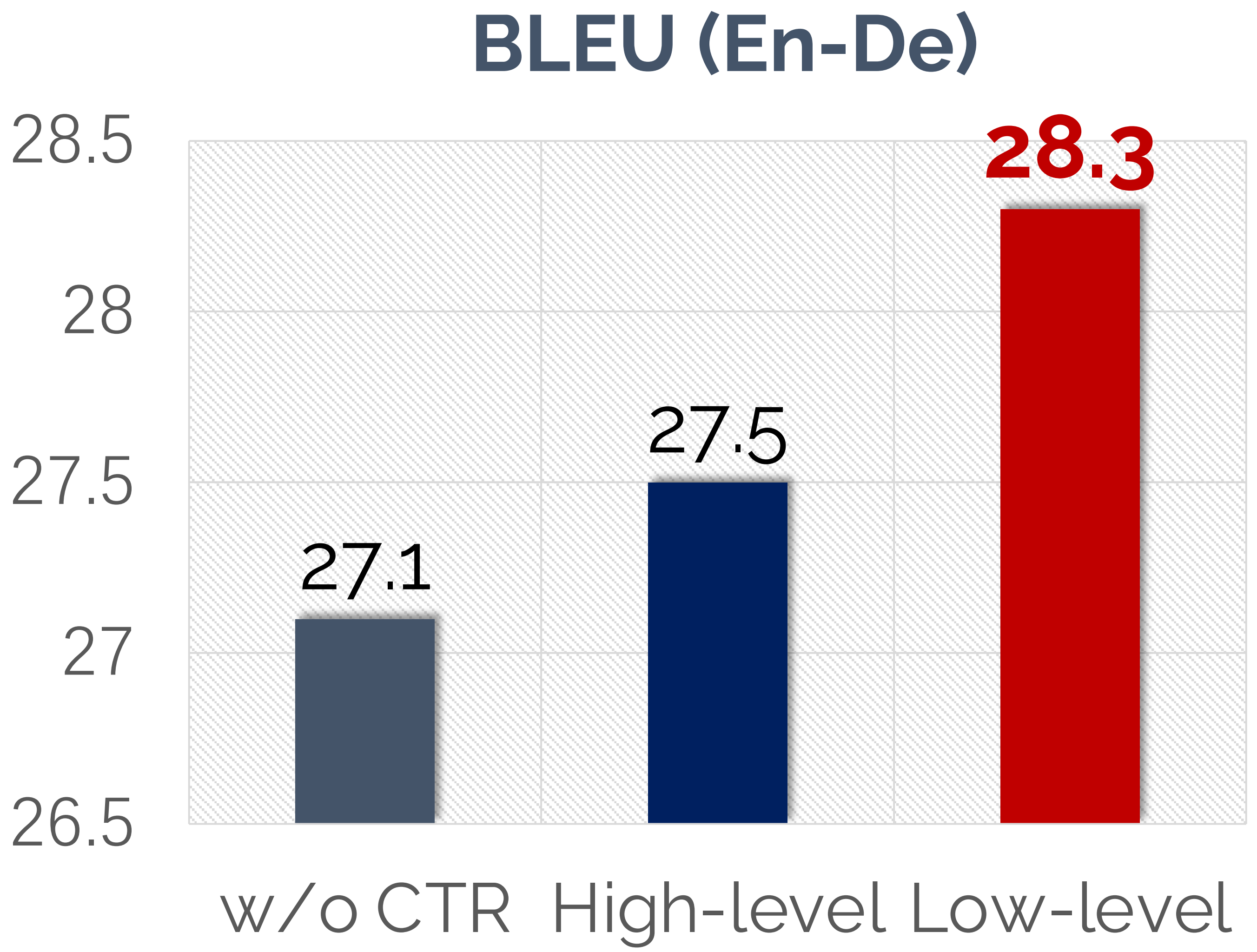
- **CTC loss**
  - Widely used in speech related tasks
  - Modeling alignment
- **L2 loss**
  - Knowledge Distillation



*: not significant

Which layer to contrast on?
——**Low level** is preferred.





🤔 Why does ConST works?

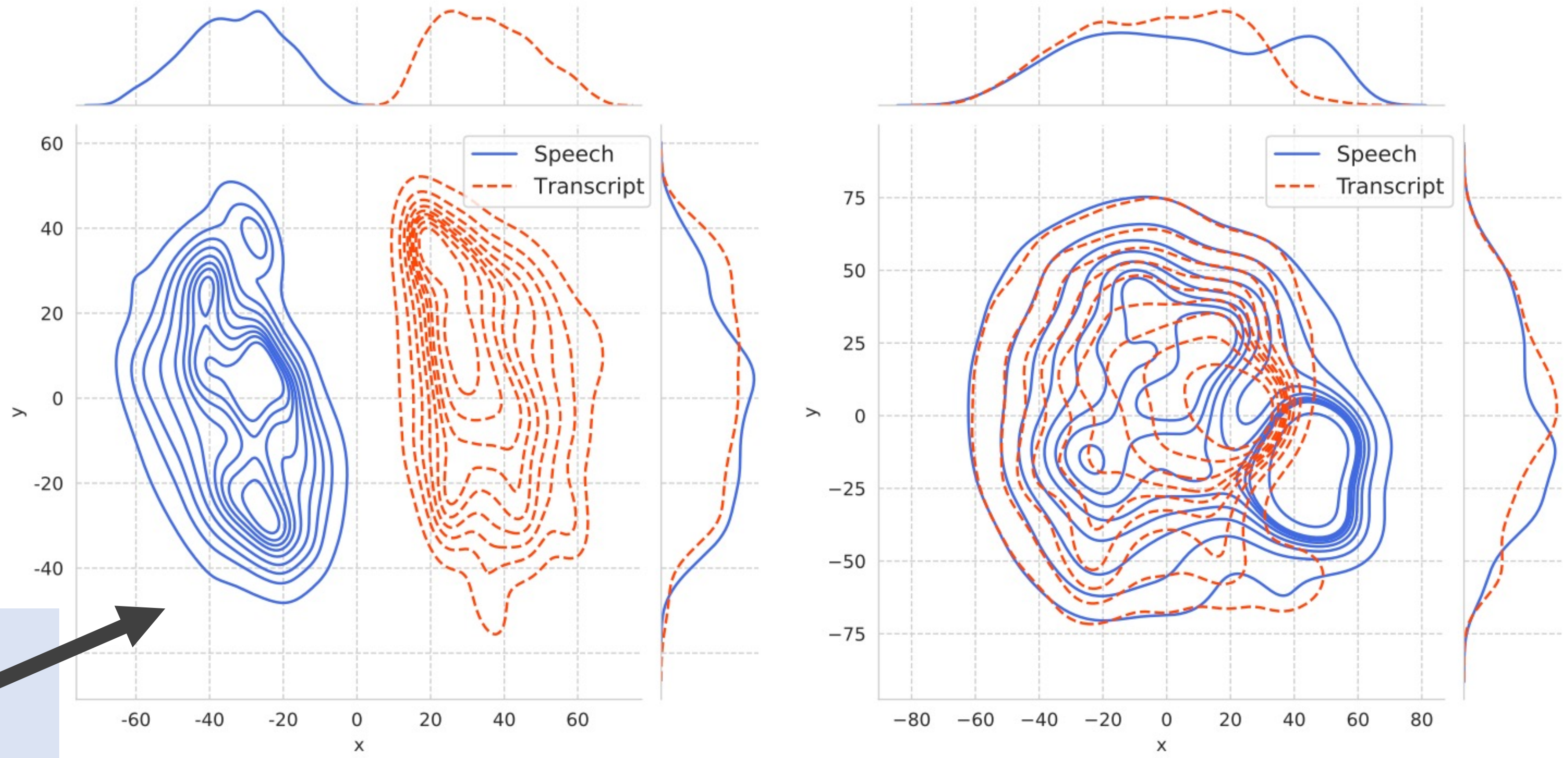
1. **Visualize** the audio and textual representation!

2. **Quantitative analysis**: A retrieval experiment.

Visualization:
CL draws the distance of two modalities!

Kernel Density Estimation (KDE) plot on "low-level" representations

XSTNet[1]: (BLEU=27.1)

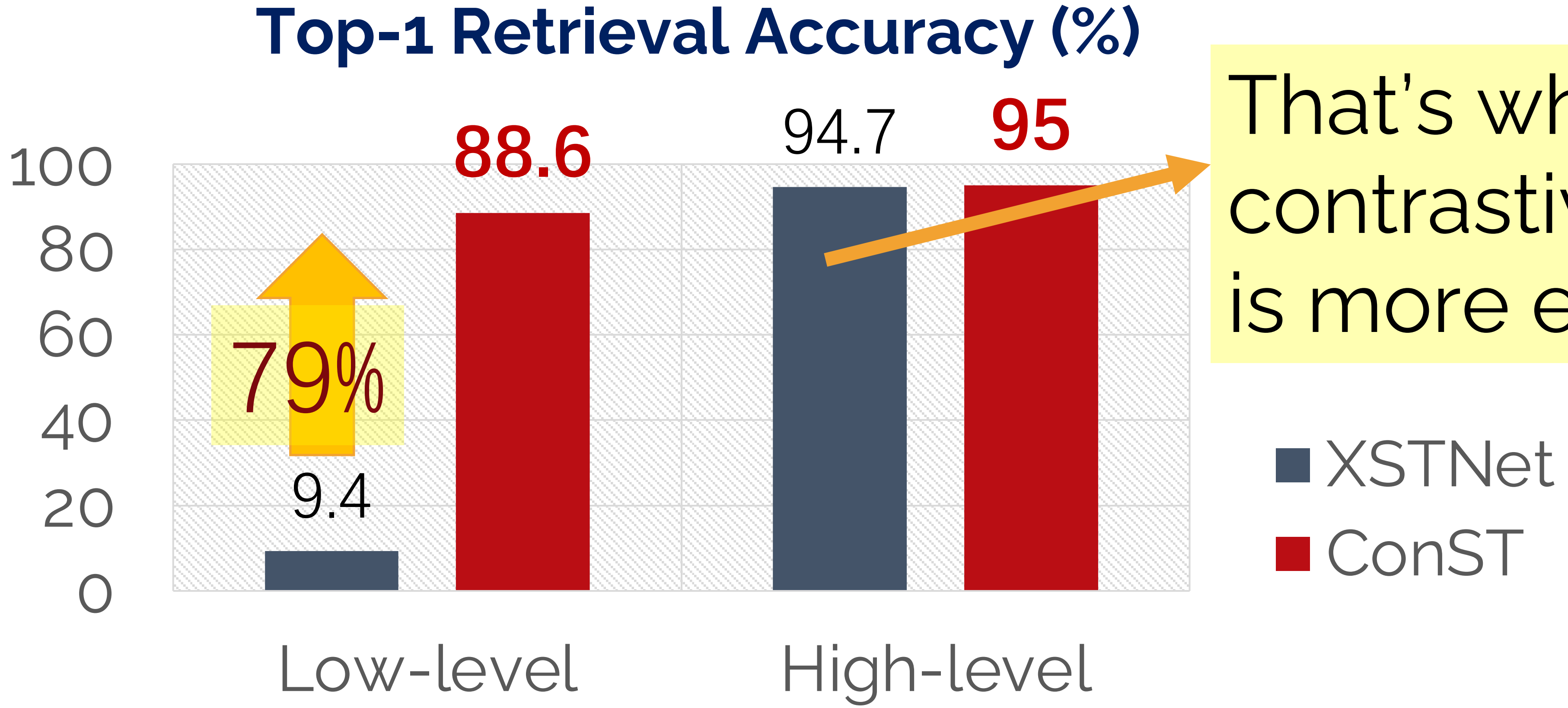


(a) w/o CTR loss (b) ConST

[1] Rong Ye, Mingxuan Wang, and Lei Li. XSTNet: End-to-end Speech Translation via Cross-modal Progressive Training. InterSpeech 2021.

Contrastively trained embedding leads to better cross-modal retrieval

- **Method**: Find the nearest (smallest cosine similarity) text based on the speech representations (low & high level)

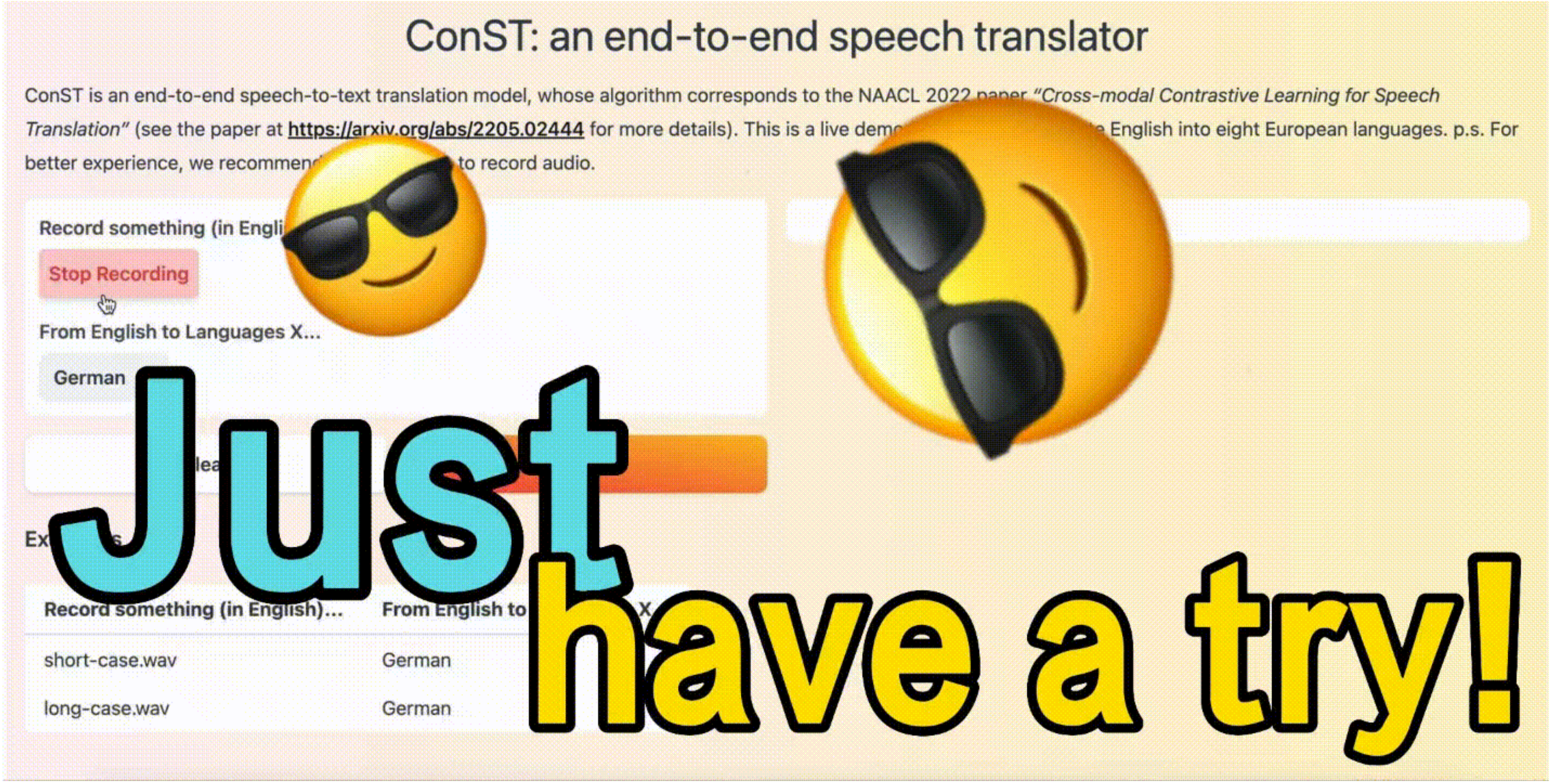


That's why **low-level** contrastive learning is more effective !

Wanna have a try?

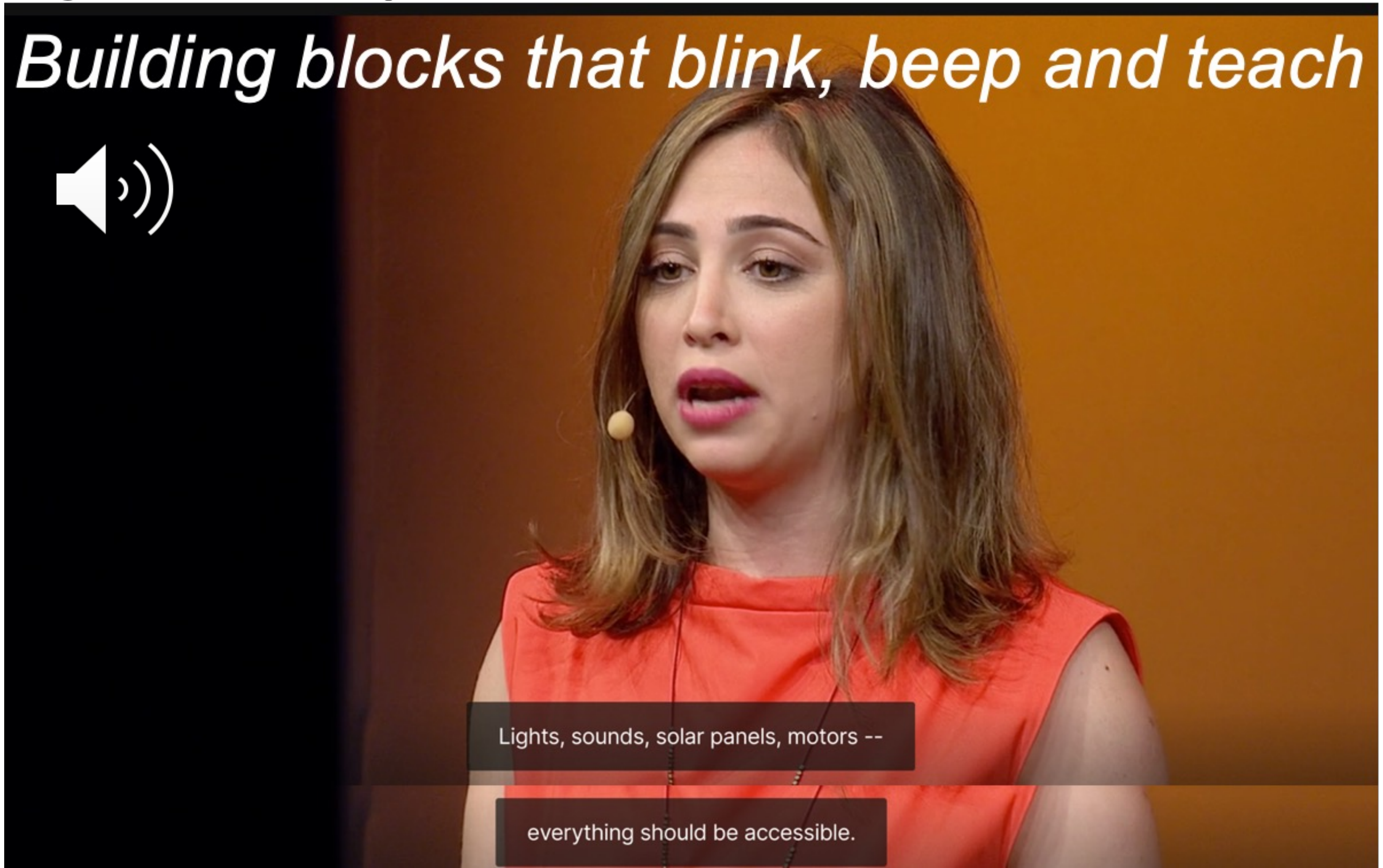
- https://huggingface.co/spaces/ReneeYe/ConST-speech2text-translator

*Best practice on ***Chrome***



Cases 1: End-to-end model avoid error propagation

Ayah Bdeir | TED 2012



Cases 1: End-to-end model avoid error propagation

- **Cascade:**

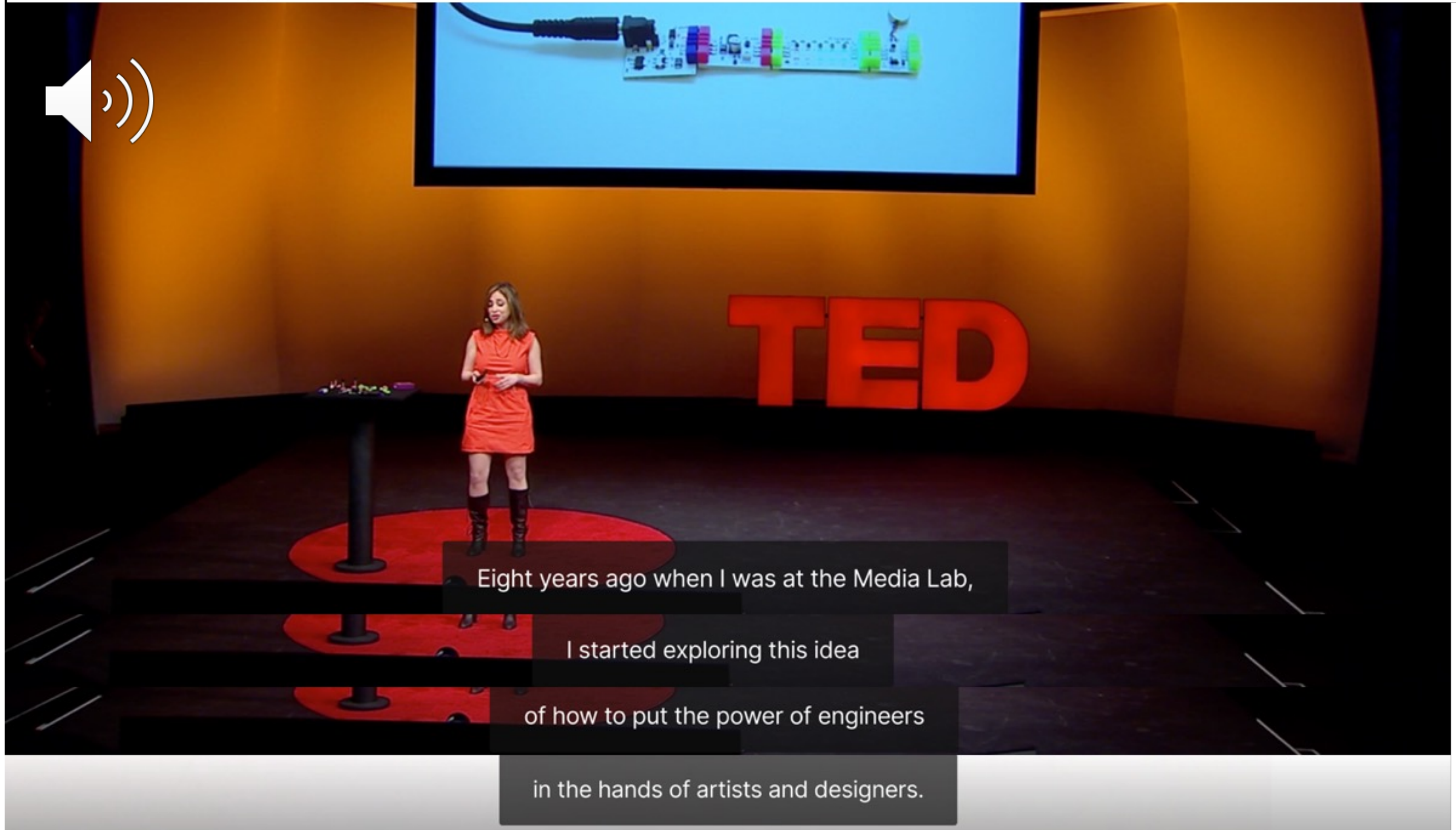
😣 **klingt** is a verb, means “sound like”

  - ❌ Licht **klingt** Solarpaneele, Motoren; alles sollte zugänglich sein.
  - Lights sounds solar panels motors everything should be accessible.
- **ConST: (correct)**
  - ✅ Licht, Geräusche, Solarpanele, Motoren, alles sollte zugänglich sein.

Case 2: Better quality then XSTNet

Ayah Bdeir | TED 2012

Building blocks that blink, beep and teach



Case 2: Better quality then XSTNet

Eight years ago when I was at the Media Lab, **I started exploring this idea of** how to …

- **XSTNet**: 😅 missing the translation on "**the idea**"
  Vor acht Jahren, als ich im Media Lab war, **begann ich zu erforschen**, wie man die …

- **ConST**:
  - ✅ Vor acht Jahren, als ich im Media Lab war, **begann ich, diese Idee zu erforschen**, wie man die …

Take-away of ConST

- Motivation of contrastive learning is to **bridge the sentence-level cross-modal representation gap.**
- **ConST**: Simple method, good performances.

- From experiments:
  - CL > CTC > L2 = simple MLT
  - **Low-level** representation is preferred to contrast on.

😉 Thanks

Paper   Code

- **Paper**: https://arxiv.org/abs/2205.02444
- **Code**: https://github.com/ReneeYe/ConST
- **E-mail**: yerong@bytedance.com