Wed 11 Dec 11 a.m. PST — 2 p.m. PST
East Exhibit Hall A-C #4937

⭐https://github.com/YUCHEN005/STAR-Adapt

NEURAL INFORMATION PROCESSING SYSTEMS

Self-Taught Recognizer: Toward Unsupervised Adaptation for Speech Foundation Models

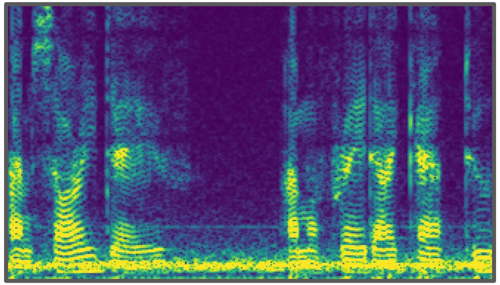
Yuchen Hu, *Chen Chen*, ***Chao-Han Huck Yang***,
Chengwei Qin, Pin-Yu Chen, Eng Siong Chng, Chao Zhang

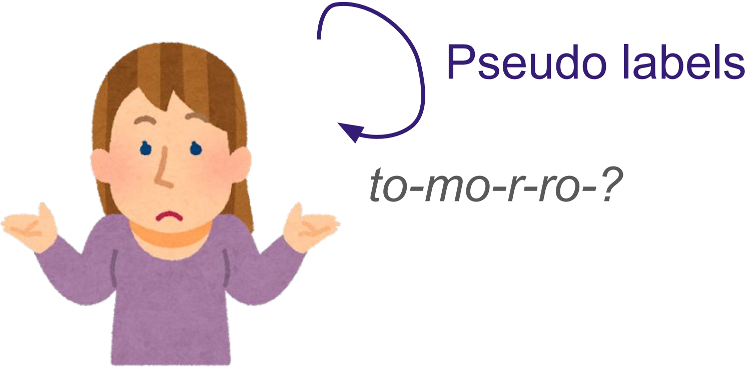
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NVIDIA

TSINGHUA UNIVERSITY 1911

How much “self-taught” processes in recognition?





Self-Taught Recognizer?

In this work …

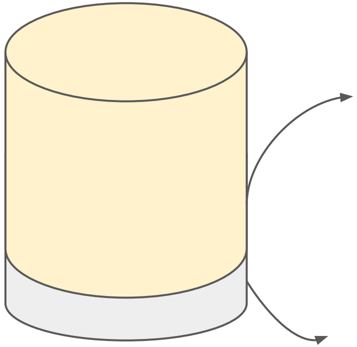
- Could we model this **“Self-Taught”** process for voice understanding?
  - Robust Automatic Speech Recognition (ASR)
  - Speech Translation (ST)

- How little labeled data we need if we could indicate
  - Confidence scores
  - Attentive scores

Task: Unsupervised Adaptation for Speech-to-Text Decoding

- ○ C: Confidence scores
- ○ A: Attentive scores

Iteratively “self-taught” label



Unlabeled transcription (90%)

Labeled transcription

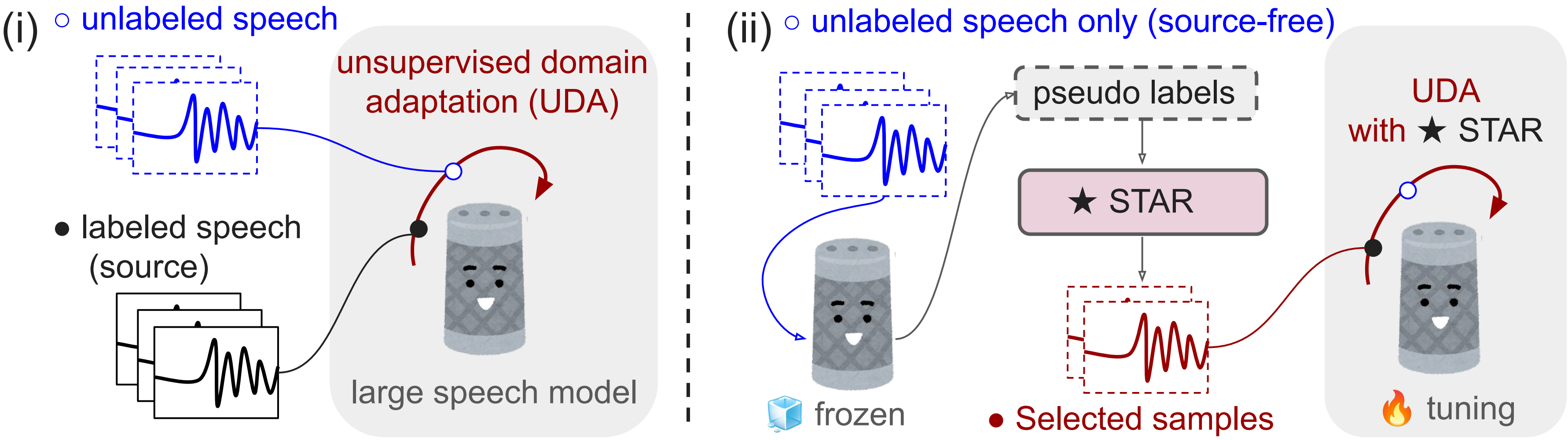
Our Contributions in this Work

1. We direct our focus on source-free UDA in **ASR & ST**, where only a pre-trained model and unlabeled speeches are required to adapt to specific target domains.

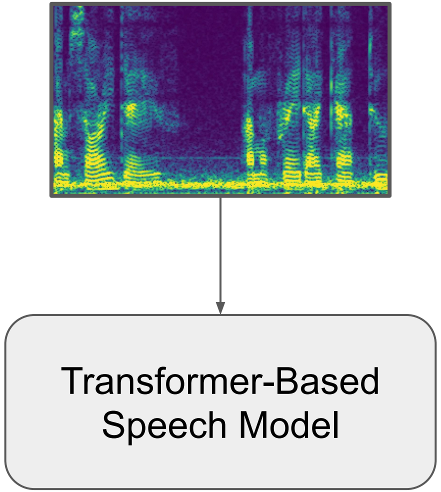
2. We introduce a self-training approach called STAR that includes a new **indicator** to evaluate the **pseudo-label quality** and achieve informed fine-tuning,

3. STAR effectively avoids the common **catastrophic forgetting problem** in adaptation.

I. Introduce Self-Taught Recognizer (STAR)



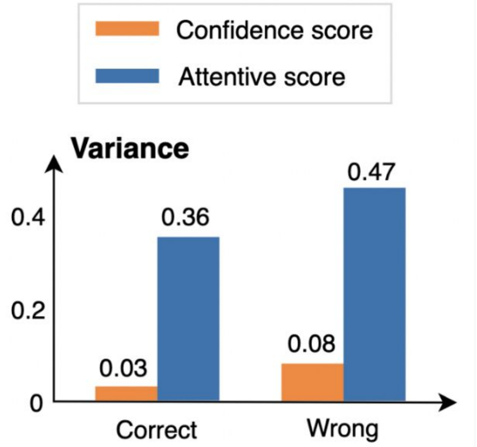
II. Confidence & Attentive Scores in Speech-to-Text Decoding (1/2)

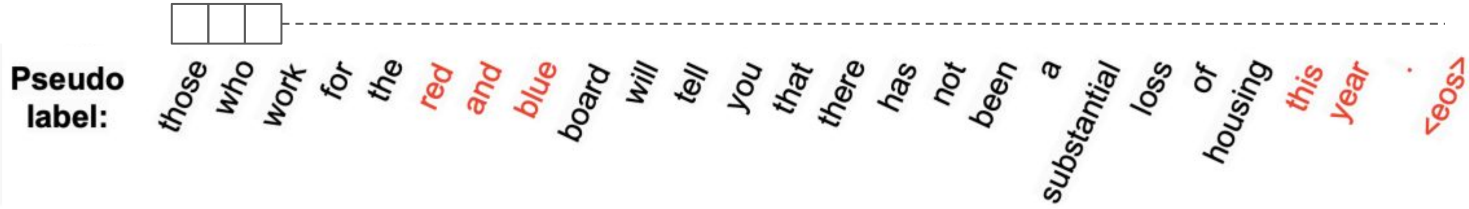


Transformer-Based Speech Model

"⟨|en|⟩⟨|transcribe|⟩⟨|notimestamps|⟩"

- ○ C: Confidence scores
- ○ A: Attentive scores

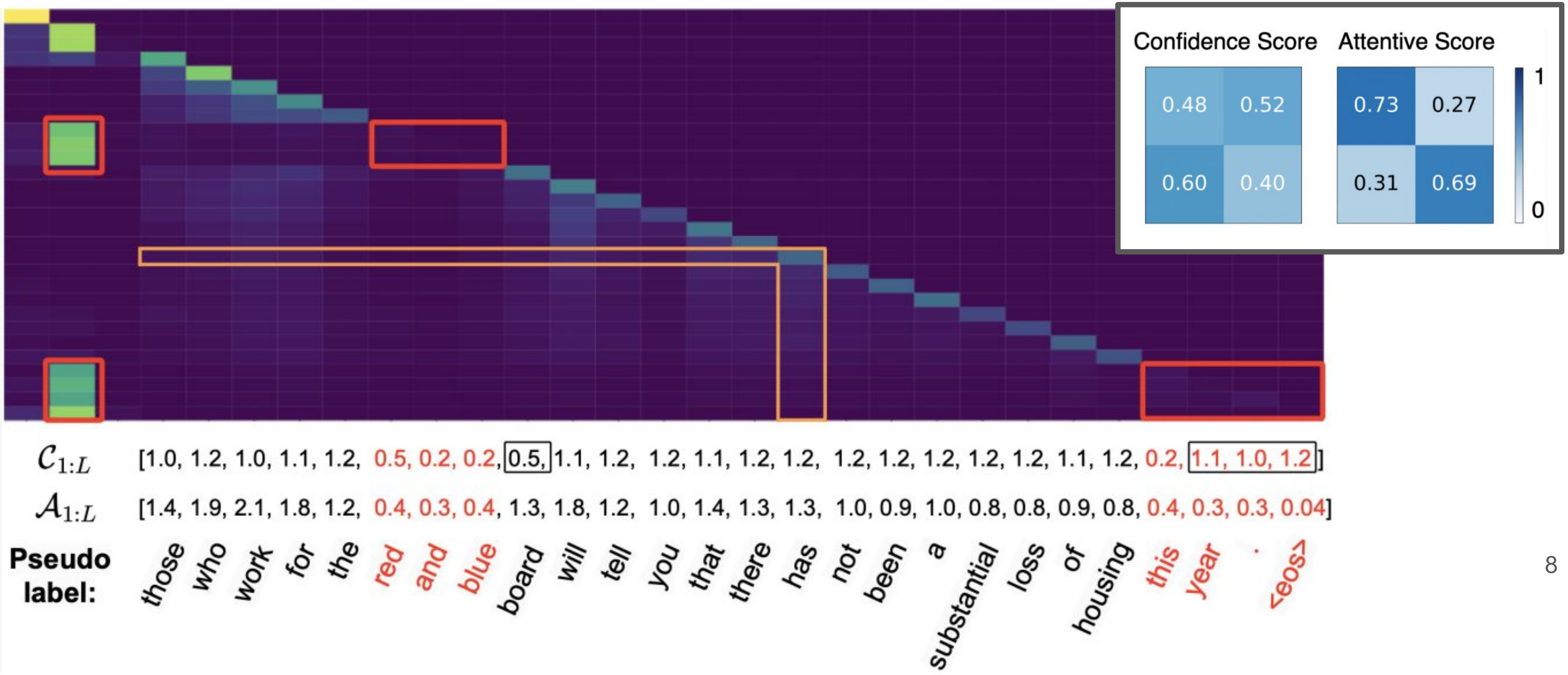




Pseudo label: those who work for the red and blue board will tell you that there has not been a substantial loss of housing this year . <eos>

II. Confidence & Attentive Scores in Speech-to-Text Decoding (2/2)

○ C: Confidence scores ○ A: Attentive scores "⟨|en|⟩⟨|transcribe|⟩⟨|notimestamps|⟩"



$\mathcal{C}_{1:L}$ [1.0, 1.2, 1.0, 1.1, 1.2, 0.5, 0.2, 0.2, 0.5, 1.1, 1.2, 1.2, 1.1, 1.2, 1.2, 1.2, 1.2, 1.2, 1.2, 1.2, 1.1, 1.2, 0.2, 1.1, 1.0, 1.2]

$\mathcal{A}_{1:L}$ [1.4, 1.9, 2.1, 1.8, 1.2, 0.4, 0.3, 0.4, 1.3, 1.8, 1.2, 1.0, 1.4, 1.3, 1.3, 1.0, 0.9, 1.0, 0.8, 0.8, 0.9, 0.8, 0.4, 0.3, 0.3, 0.04]

Pseudo label: those who work for the red and blue board will tell you that there has not been a substantial loss of housing this year . <eos>

II. STAR Indicator: Reliable and Stable (1/2)

meta-thresholds

$$\mathcal{S}_l^{\text{conf}} = [\sigma(\mathcal{A}_l^2/\mathcal{C}_l - \lambda) + \sigma(\mathcal{C}_l^2/\mathcal{A}_l - \lambda)] * \mathcal{A}_l,$$

activation functions

focal loss style smooth

$$\mathcal{S}_l^{\text{cons}} = [\sigma(\lambda - \mathcal{A}_l^2/\mathcal{C}_l) * \sigma(\lambda - \mathcal{C}_l^2/\mathcal{A}_l)] \; * \mathcal{A}_l * e^{(\mathcal{C}_l - \mathcal{A}_l)/\tau}.$$

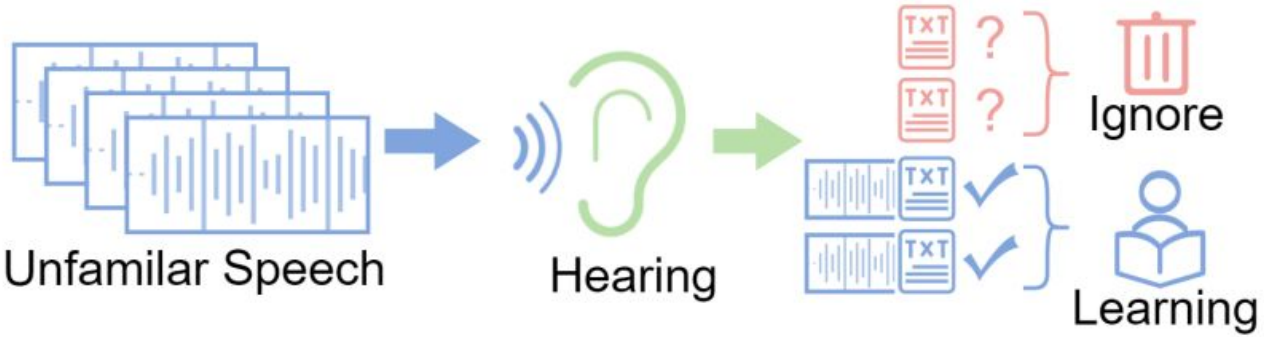
- C: Confidence scores
- A: Attentive scores

II. STAR Indicator: Reliable and Stable (2/2)

$$\widetilde{\mathcal{L}}_{\mathrm{ASR}}(x,\hat{y}) = \sum_{l=1}^{L} -\log \mathcal{P}_{\theta}(\hat{y}_l|\hat{y}_{l-1:1},x) * \mathcal{S}_l; \quad \text{where } \mathcal{S}_l = \mathcal{S}_l^{\text{conf}} + \mathcal{S}_l^{\text{cons}}.$$

or LST

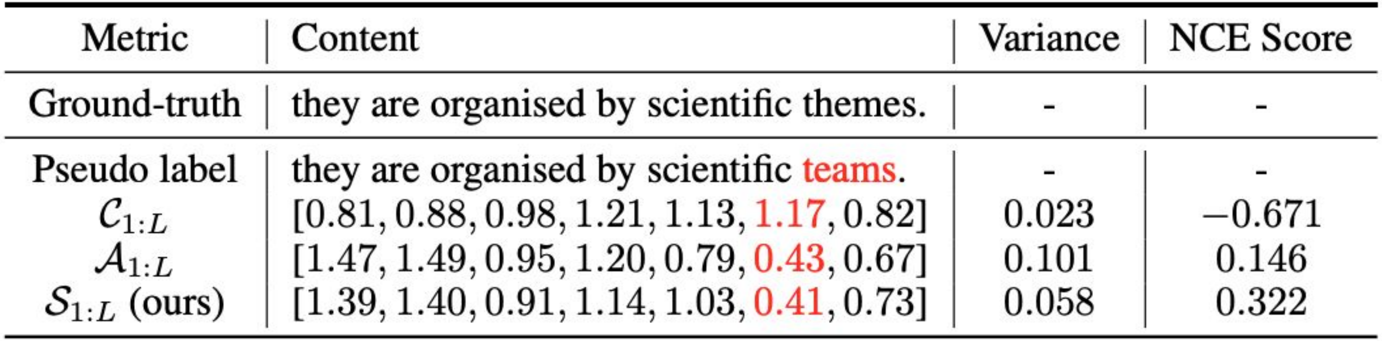
- C: Confidence scores
- A: Attentive scores



*Monte Carlo Sampling

III. STAR Case Study

Normalized cross-entropy (NCE)

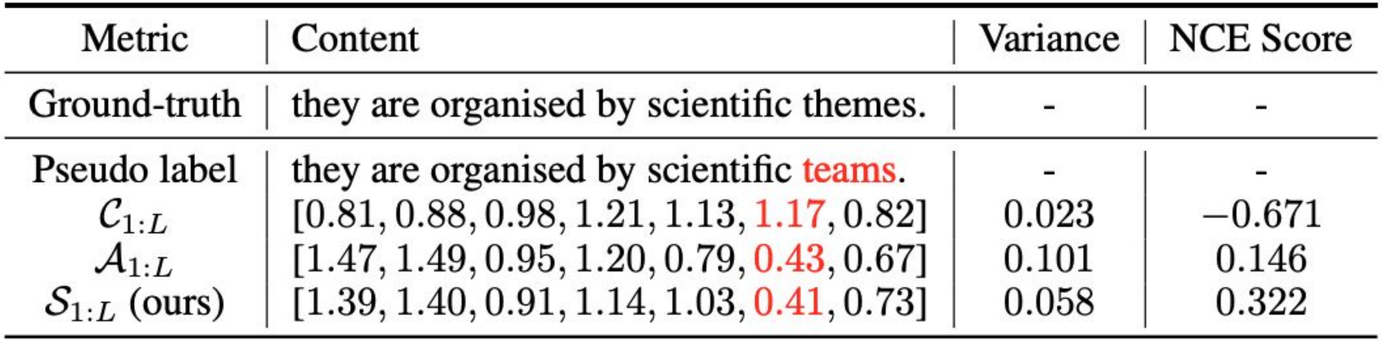


Metric	Content	Variance ↓	NCE Score ↑
Ground-truth	they are organised by scientific themes.	-	-
Pseudo label	they are organised by scientific teams.	-	-
$\mathcal{C}_{1:L}$	$[0.81, 0.88, 0.98, 1.21, 1.13, 1.17, 0.82]$	0.023	-0.671
$\mathcal{A}_{1:L}$	$[1.47, 1.49, 0.95, 1.20, 0.79, 0.43, 0.67]$	0.101	0.146
$\mathcal{S}_{1:L}$ (ours)	$[1.39, 1.40, 0.91, 1.14, 1.03, 0.41, 0.73]$	0.058	0.322

Common voice hindi accent English
ID: “en_19795319”

III. STAR Case Study

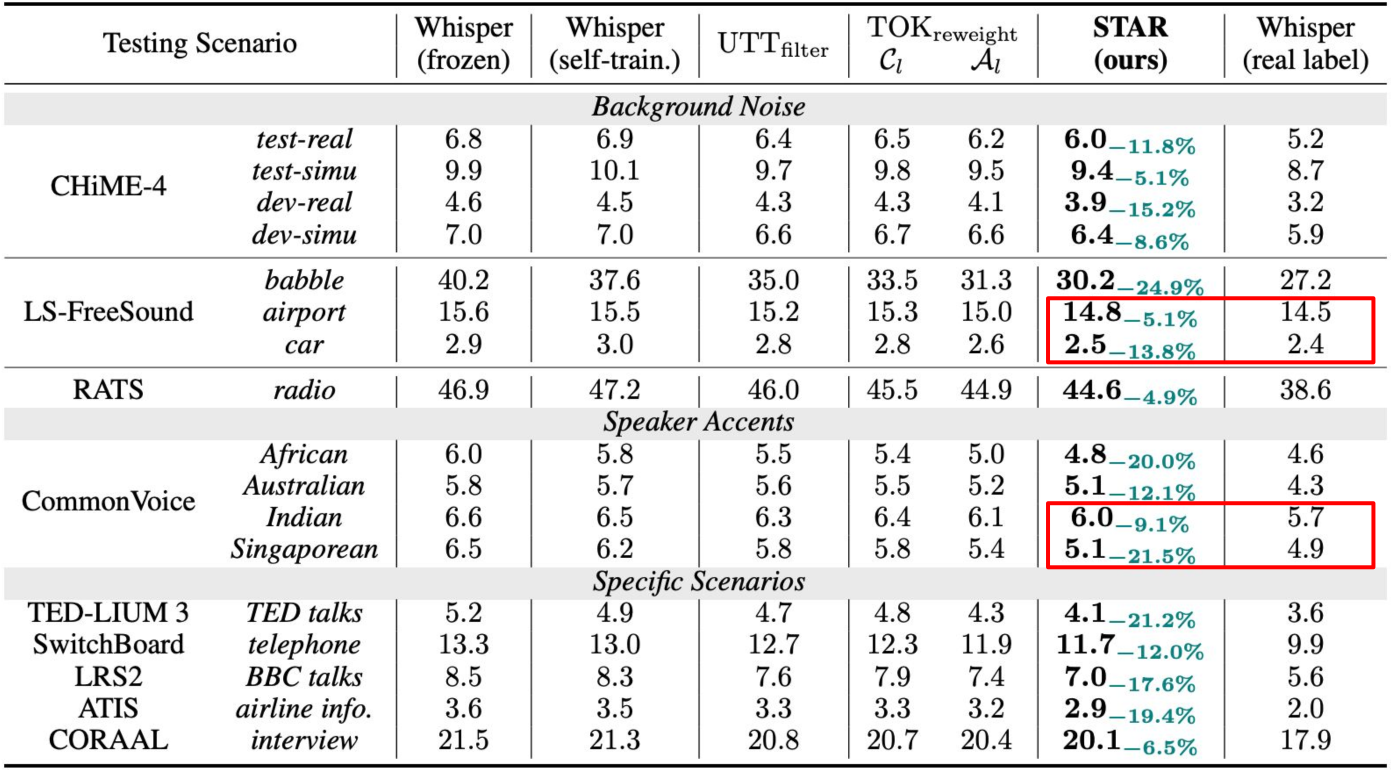
Normalized cross-entropy (NCE)



Metric	Content	Variance ↓	NCE Score ↑
Ground-truth	they are organised by scientific themes.	-	-
Pseudo label	they are organised by scientific teams.	-	-
$\mathcal{C}_{1:L}$	$[0.81, 0.88, 0.98, 1.21, 1.13, 1.17, 0.82]$	0.023	-0.671
$\mathcal{A}_{1:L}$	$[1.47, 1.49, 0.95, 1.20, 0.79, 0.43, 0.67]$	0.101	0.146
$\mathcal{S}_{1:L}$ (ours)	$[1.39, 1.40, 0.91, 1.14, 1.03, 0.41, 0.73]$	0.058	0.322

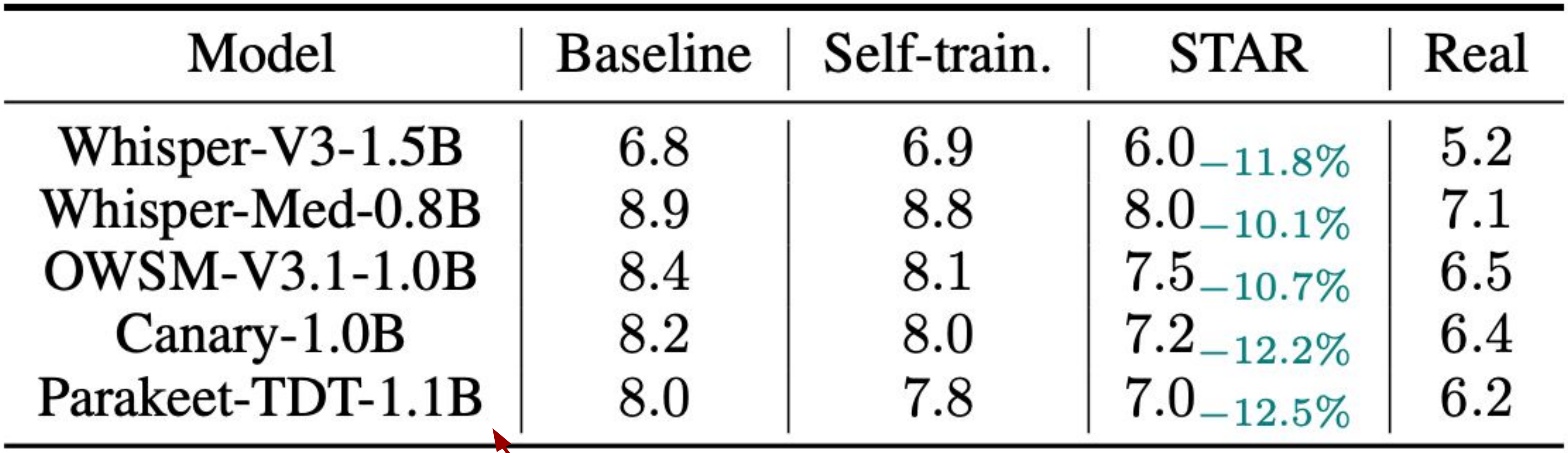
Common voice hindi accent English
ID: “en_19795319”

III. STAR Indicator is Effective cross Noisy Datasets



Testing Scenario		Whisper (frozen)	Whisper (self-train.)	UTT$_{\text{filter}}$	TOK$_{\text{reweight}}$ $\mathcal{C}_l$	TOK$_{\text{reweight}}$ $\mathcal{A}_l$	**STAR (ours)**	Whisper (real label)
Background Noise								
CHiME-4	*test-real*	6.8	6.9	6.4	6.5	6.2	**6.0**$_{-11.8\%}$	5.2
	test-simu	9.9	10.1	9.7	9.8	9.5	**9.4**$_{-5.1\%}$	8.7
	dev-real	4.6	4.5	4.3	4.3	4.1	**3.9**$_{-15.2\%}$	3.2
	dev-simu	7.0	7.0	6.6	6.7	6.6	**6.4**$_{-8.6\%}$	5.9
LS-FreeSound	*babble*	40.2	37.6	35.0	33.5	31.3	**30.2**$_{-24.9\%}$	27.2
	airport	15.6	15.5	15.2	15.3	15.0	**14.8**$_{-5.1\%}$	14.5
	car	2.9	3.0	2.8	2.8	2.6	**2.5**$_{-13.8\%}$	2.4
RATS	*radio*	46.9	47.2	46.0	45.5	44.9	**44.6**$_{-4.9\%}$	38.6
Speaker Accents								
CommonVoice	*African*	6.0	5.8	5.5	5.4	5.0	**4.8**$_{-20.0\%}$	4.6
	Australian	5.8	5.7	5.6	5.5	5.2	**5.1**$_{-12.1\%}$	4.3
	Indian	6.6	6.5	6.3	6.4	6.1	**6.0**$_{-9.1\%}$	5.7
	Singaporean	6.5	6.2	5.8	5.8	5.4	**5.1**$_{-21.5\%}$	4.9
Specific Scenarios								
TED-LIUM 3	*TED talks*	5.2	4.9	4.7	4.8	4.3	**4.1**$_{-21.2\%}$	3.6
SwitchBoard	*telephone*	13.3	13.0	12.7	12.3	11.9	**11.7**$_{-12.0\%}$	9.9
LRS2	*BBC talks*	8.5	8.3	7.6	7.9	7.4	**7.0**$_{-17.6\%}$	5.6
ATIS	*airline info.*	3.6	3.5	3.3	3.3	3.2	**2.9**$_{-19.4\%}$	2.0
CORAAL	*interview*	21.5	21.3	20.8	20.7	20.4	**20.1**$_{-6.5\%}$	17.9

III. STAR Indicator works for both Whisper & RNN-T



Model	Baseline	Self-train.	STAR	Real
Whisper-V3-1.5B	6.8	6.9	$6.0_{-11.8\%}$	5.2
Whisper-Med-0.8B	8.9	8.8	$8.0_{-10.1\%}$	7.1
OWSM-V3.1-1.0B	8.4	8.1	$7.5_{-10.7\%}$	6.5
Canary-1.0B	8.2	8.0	$7.2_{-12.2\%}$	6.4
Parakeet-TDT-1.1B	8.0	7.8	$7.0_{-12.5\%}$	6.2

Conformer RNN-T

III. STAR Indicator is Sample-Efficient

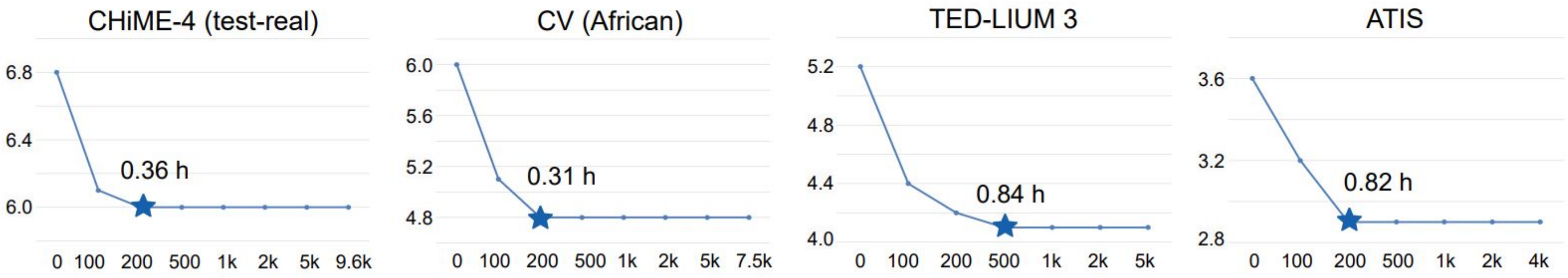
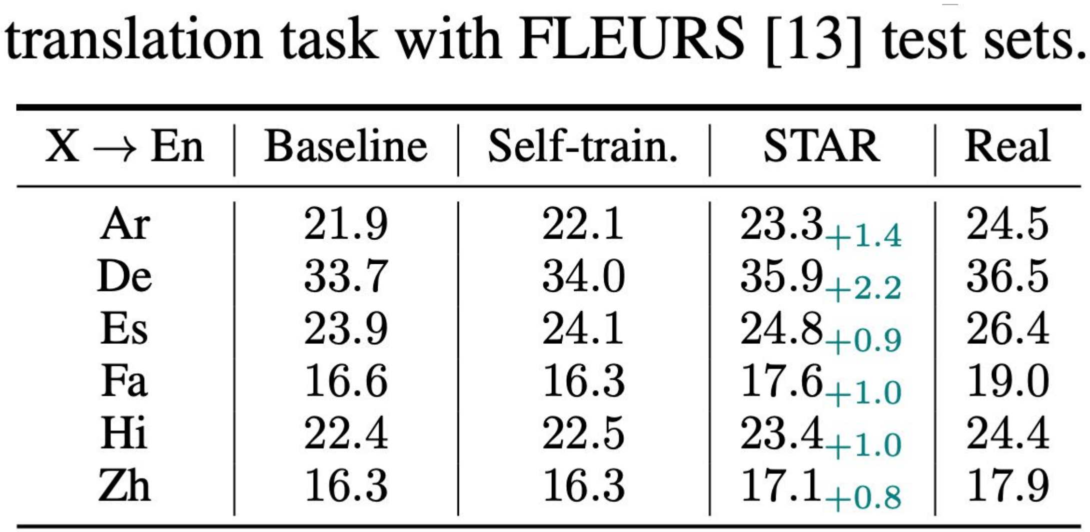


Figure 3: WER (%) results with different numbers of unlabeled training samples. The minimum required data amount (in hours) to obtain the best performance is highlighted in the star mark.

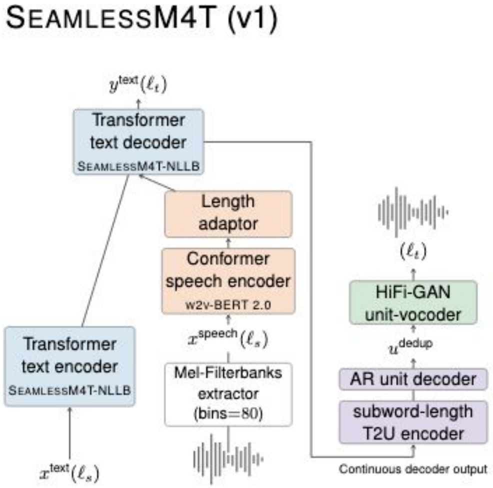
III. STAR works for Multilingual Speech Translation Tasks

Backbone 2.3B Model - SeamlessM4T (Meta)

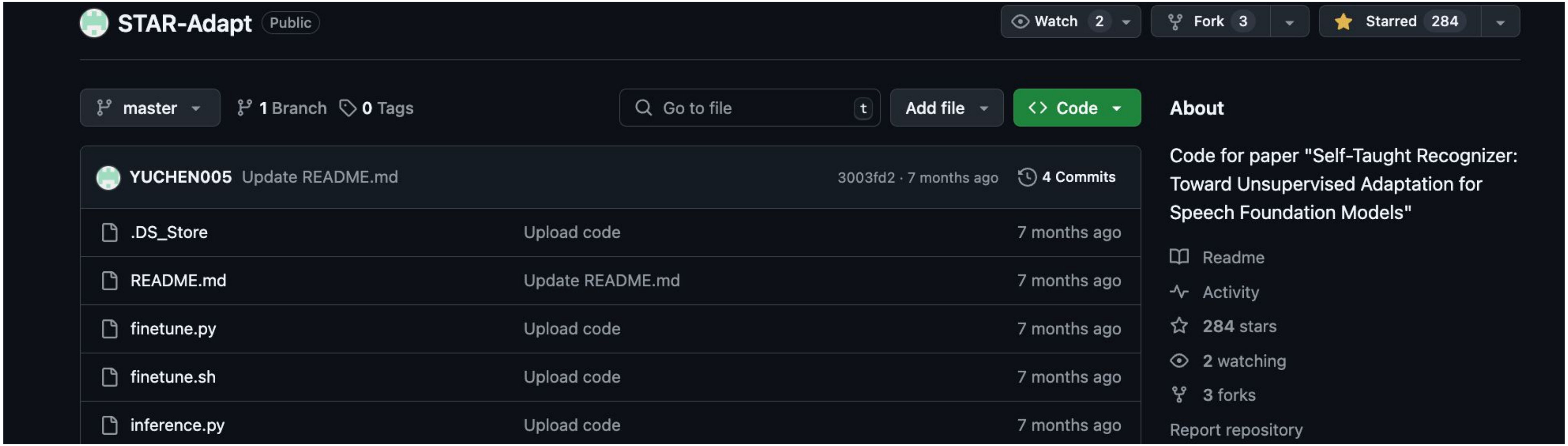


translation task with FLEURS [13] test sets.

X → En	Baseline	Self-train.	STAR	Real
Ar	21.9	22.1	$23.3_{+1.4}$	24.5
De	33.7	34.0	$35.9_{+2.2}$	36.5
Es	23.9	24.1	$24.8_{+0.9}$	26.4
Fa	16.6	16.3	$17.6_{+1.0}$	19.0
Hi	22.4	22.5	$23.4_{+1.0}$	24.4
Zh	16.3	16.3	$17.1_{+0.8}$	17.9



Acknowledgement



YUCHEN005	Update README.md	3003fd2 · 7 months ago · 4 Commits
.DS_Store	Upload code	7 months ago
README.md	Update README.md	7 months ago
finetune.py	Upload code	7 months ago
finetune.sh	Upload code	7 months ago
inference.py	Upload code	7 months ago

Code for paper "Self-Taught Recognizer: Toward Unsupervised Adaptation for Speech Foundation Models"

⭐https://github.com/YUCHEN005/STAR-Adapt

- **{yuchen005, chen1436}@e.ntu.edu.sg and hucky@nvidia.com**