# Losses Can Be Blessings: Routing Self-Supervised Speech Representations Towards Efficient Multilingual and Multitask Speech Processing

*NeurIPS 2022*

***Yonggan Fu***, Yang Zhang, Kaizhi Qian, Zhifan Ye,

Zhongzhi Yu, Cheng-I Lai, Yingyan (Celine) Lin

# Motivation: Demanding ASR Systems

- **A growing demand**: Deploy DNN-based Automatic Speech Recognition (ASR) systems on mobile devices

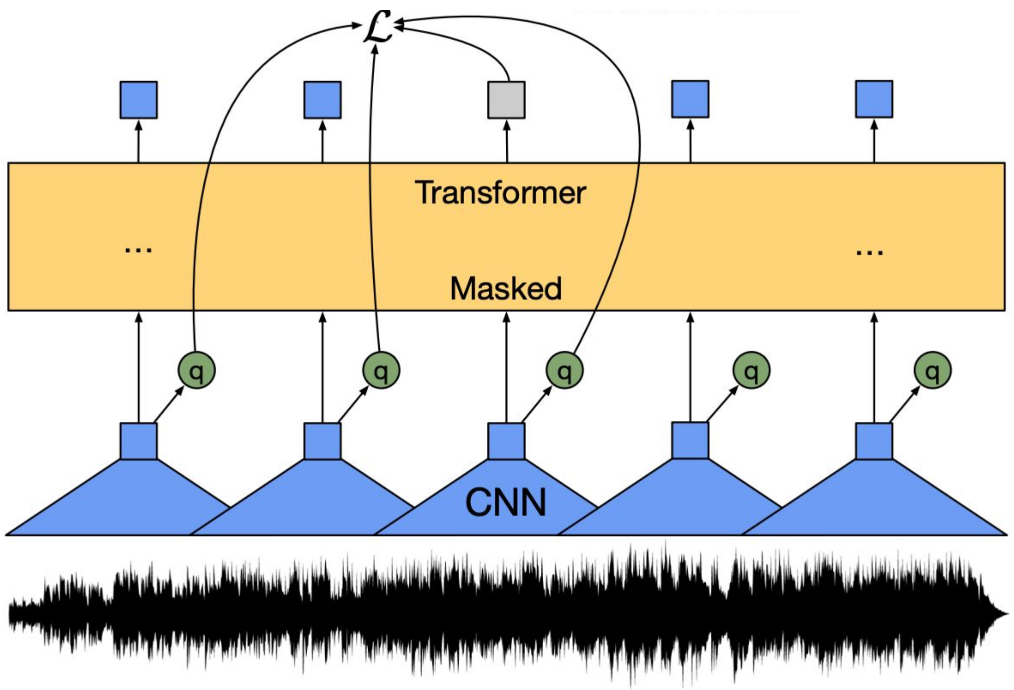
# Motivation: Demanding ASR Systems

- **A growing demand**: Deploy DNN-based Automatic Speech Recognition (ASR) systems on mobile devices

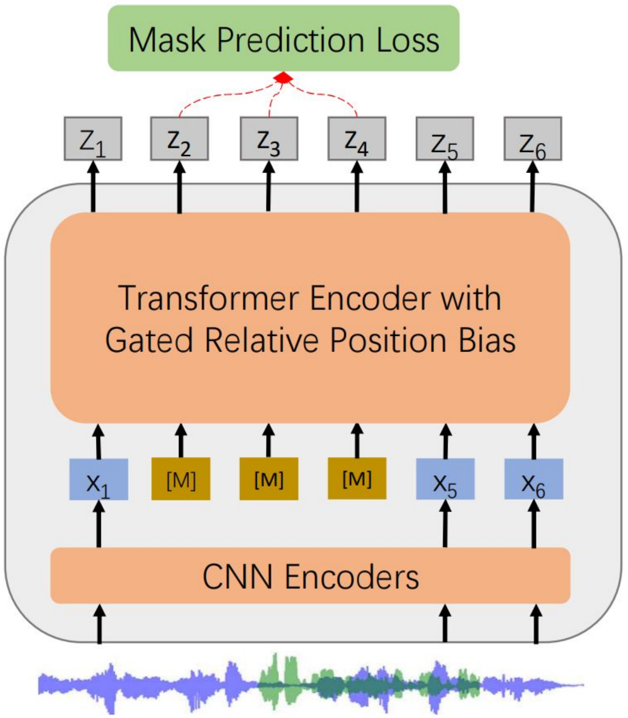
***Challenge:*** **The big data regime is not always possible for low-resource spoken languages**

# Speech SSL Models: Enable Low-resource ASR

😀 **SOTA low-resource ASR solutions**: Self-supervised learning (SSL) towards rich speech representations



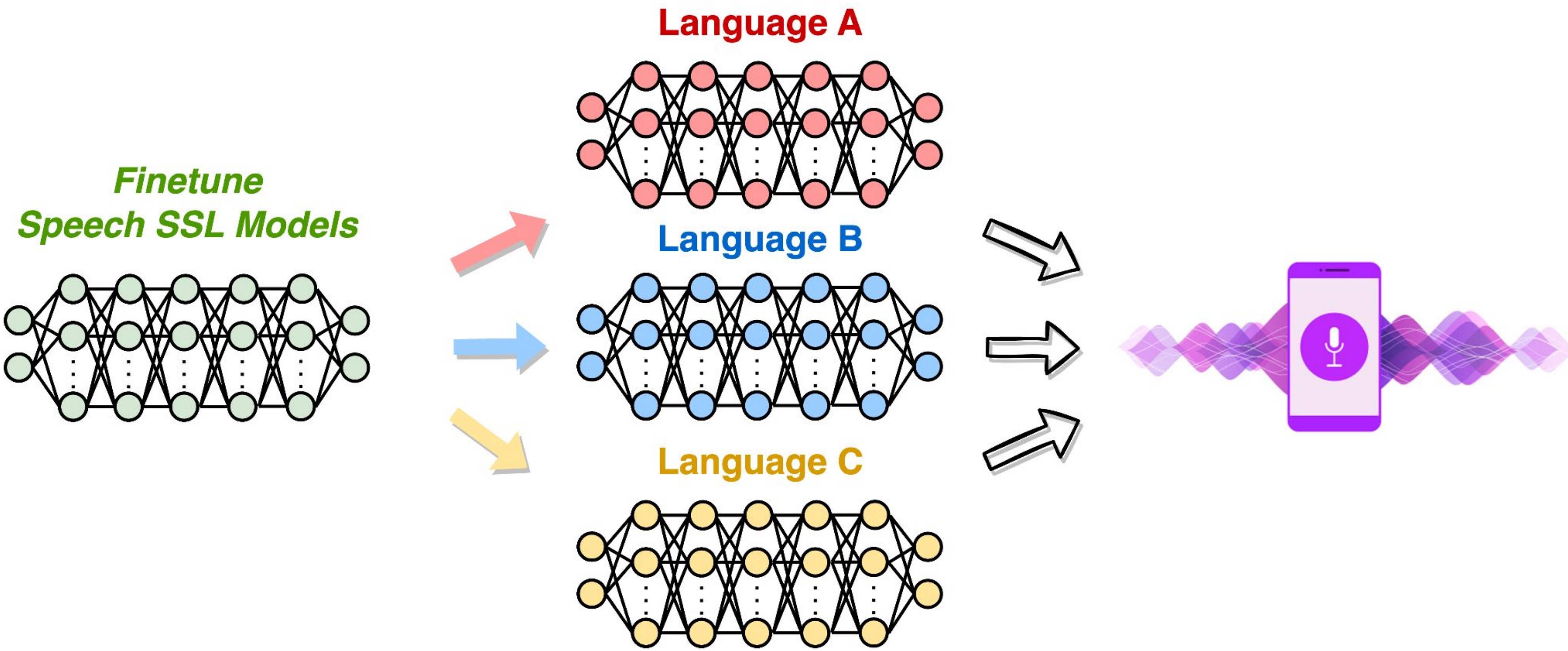

**Wav2vec 2.0 [*NeurIPS'20*]**




**WavLM [*JSTSP'22*]**

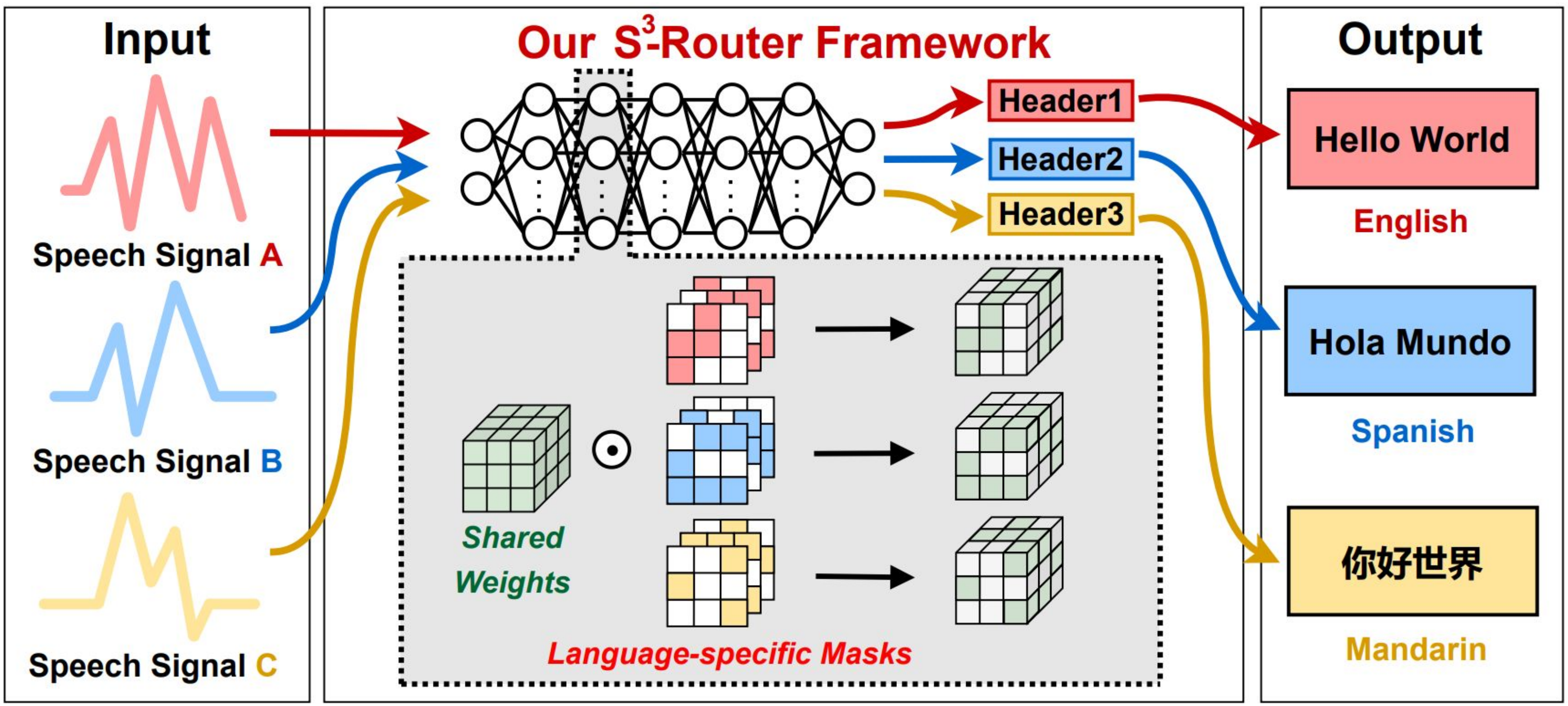
# Speech SSL Models: Efficiency Concerns

- ☹ **Prohibitive complexity of speech SSL models**
  - Especially for multilingual/multitask speech processing due to the pretrain-then-finetune paradigm



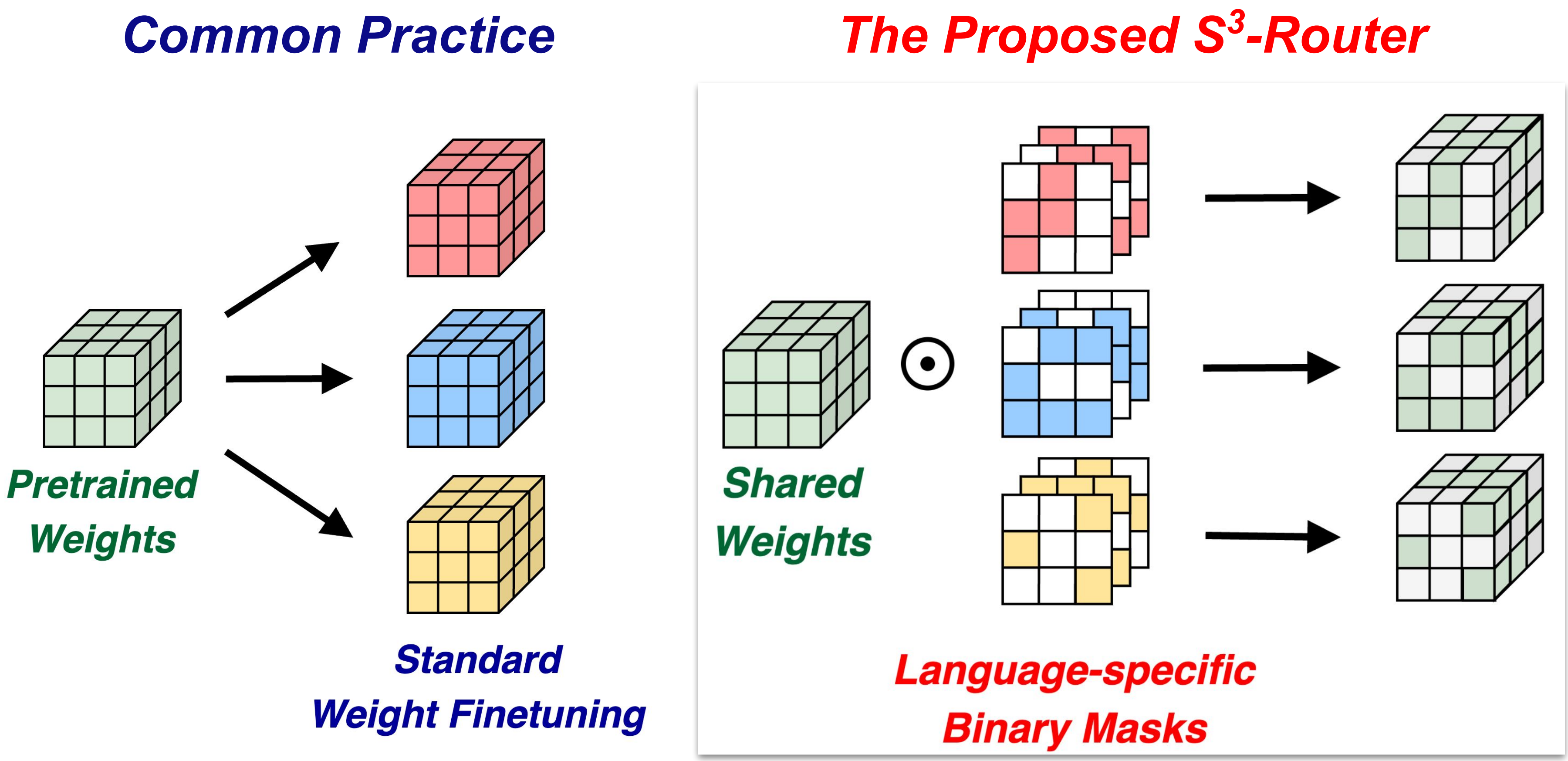
# Our Proposed $S^3$-Router Framework

- **Key idea: S**elf-**S**upervised **S**peech Representation Router
  - Finetune model connections on top of shared weights via optimizing language-/task-specific binary masks



# Our Proposed S$^3$-Router Framework

- **Key insight:** Model sparsity can be utilized to encode language-/task-specific information



# $S^3$-Router: Method Formulation

- **Formulation of binary mask optimization**
  - **Forward:** Activate only top $k_t$ elements
  - **Backward:** All elements in $m_t$ are updated via STE

$$\arg\min_{m_t} \sum_{(x_t, y_t) \in D_t} \ell_t(f(\underbrace{m_t \odot \theta_{SSL}}_{\text{Apply language-/task-specific binary masks}}, x_t), y_t) \quad s.t. \quad \underbrace{||m_t||_0 \leqslant k_t}_{\text{Sparsity constraint}}$$

$\theta_{SSL}$ : *SSL pretrained weights (fixed)*

$t$ : *The index of languages/tasks*

# $S^3$-Router: Method Formulation

- **Formulation of binary mask optimization**
  - **Forward:** Activate only top $k_t$ elements
  - **Backward:** All elements in $m_t$ are updated via STE

$$\arg\min_{m_t} \sum_{(x_t, y_t) \in D_t} \ell_t(f(\underbrace{m_t \odot \theta_{SSL}}_{\text{Apply language-/task-specific binary masks}}, x_t), y_t) \quad s.t. \quad \underbrace{||m_t||_0 \leqslant k_t}_{\text{Sparsity constraint}}$$

***How to initialize the binary masks?***

# S$^3$-Router: Method Formulation

- **Mask initialization**
  - ☹ Random initialization: ***No prior is utilized***
  - ☹ Weight magnitude based initialization: ***Poor trainability***

# S$^3$-Router: Method Formulation

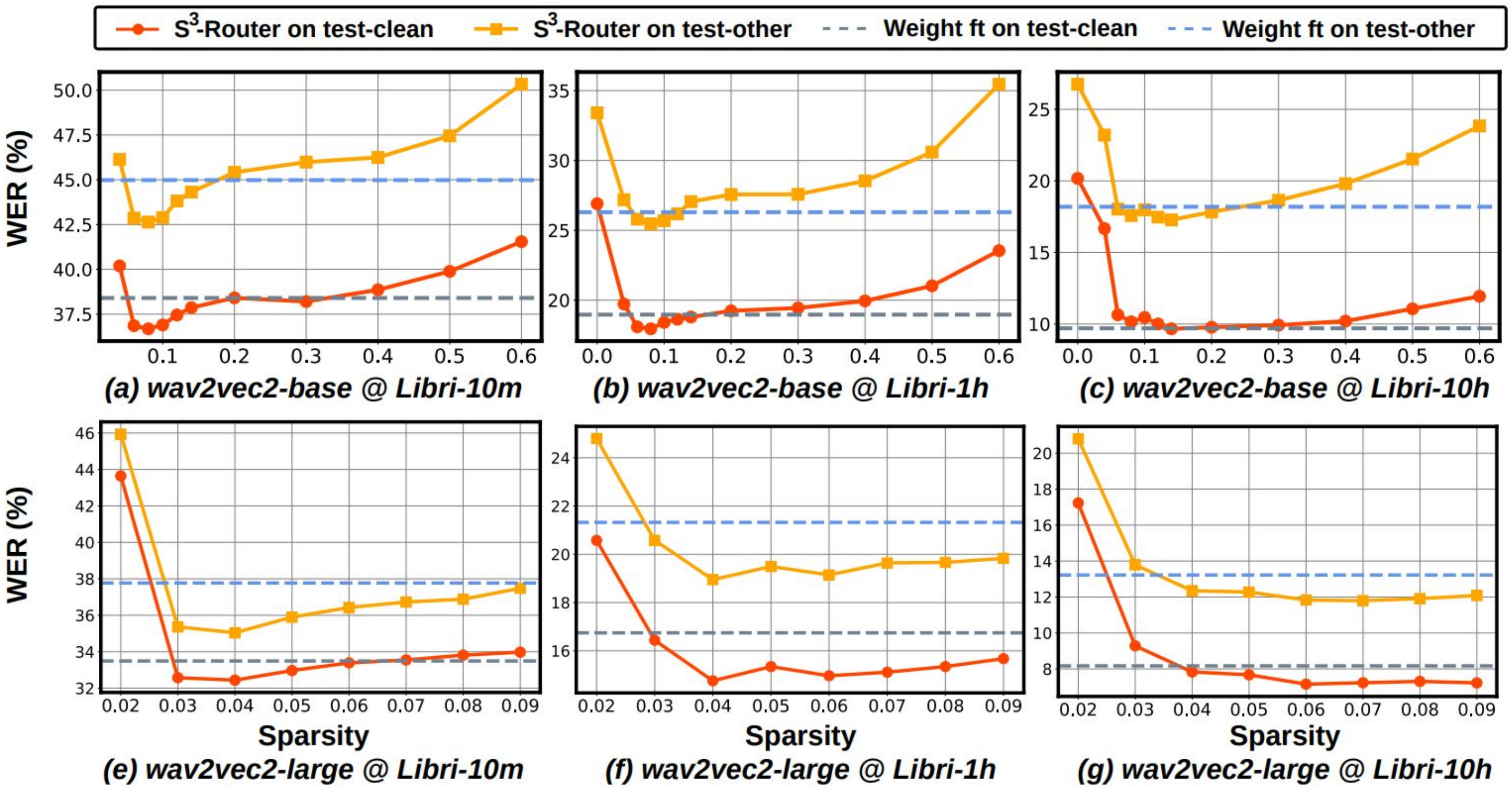
- **Mask initialization**
  - ☹ Random initialization: ***No prior is utilized***
  - ☹ Weight magnitude based initialization: ***Poor trainability***

⇩

  - 😀 **Our Proposed Order-Preserving Random Initialization**
    - Random mask values for boosted trainability
    - Maintain the orders of weight magnitudes as priors

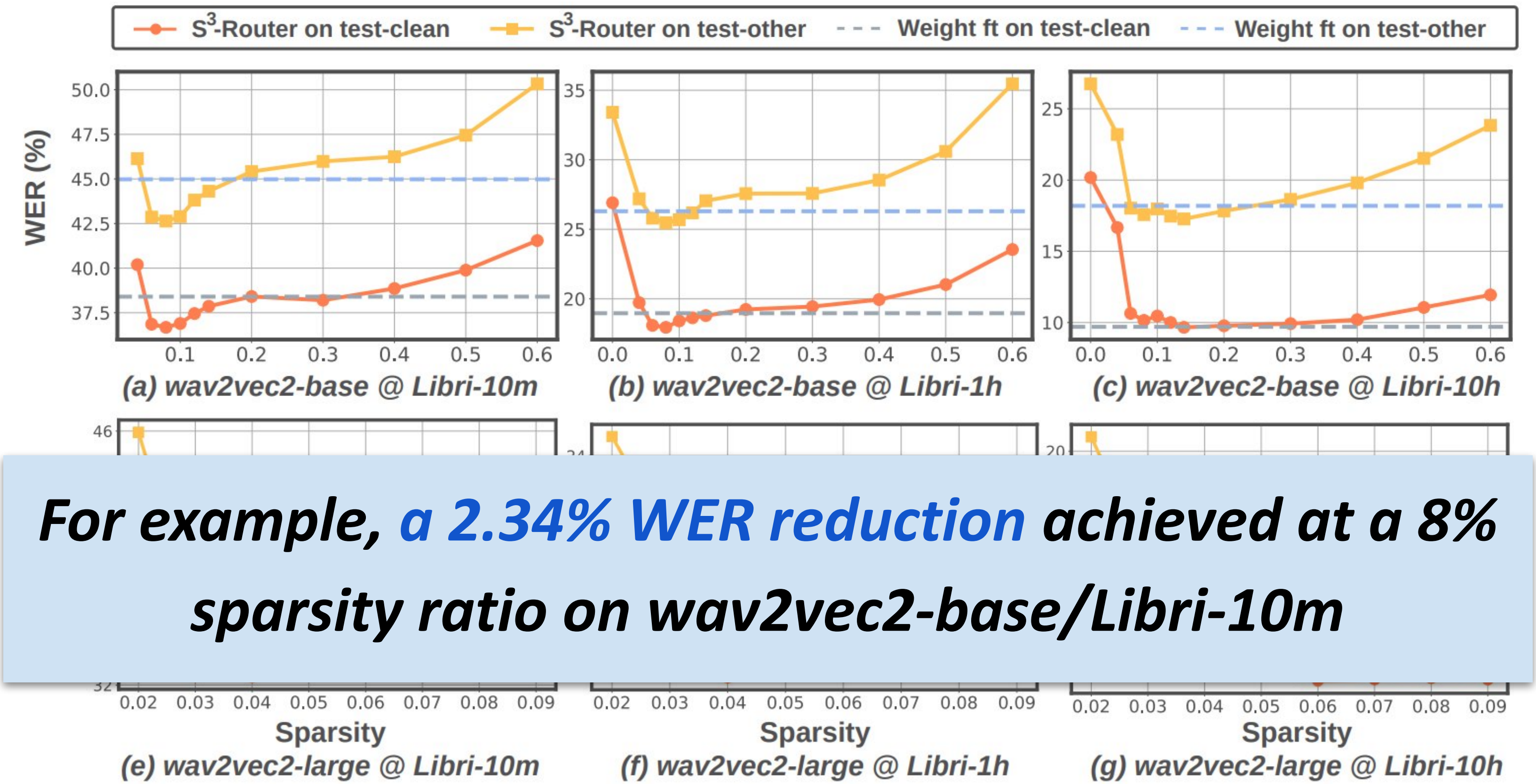
# $S^3$-Router's App. 1: A New Finetuning Paradigm

- **Discarding ≤10% weights is all you need**
  - Consistently outperform the standard weight finetuning in terms of the achievable word error rate (WER)



# S$^3$-Router's App. 1: A New Finetuning Paradigm

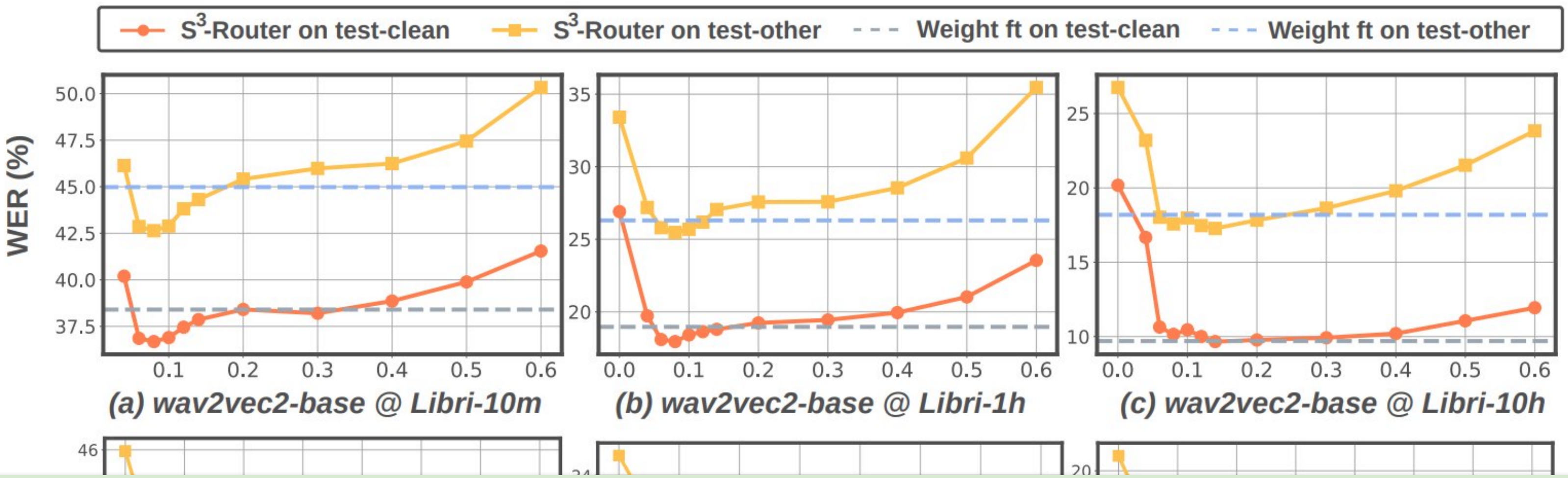
- **Discarding ≤10% weights is all you need**
  - Consistently outperform the standard weight finetuning in terms of the achievable word error rate (WER)



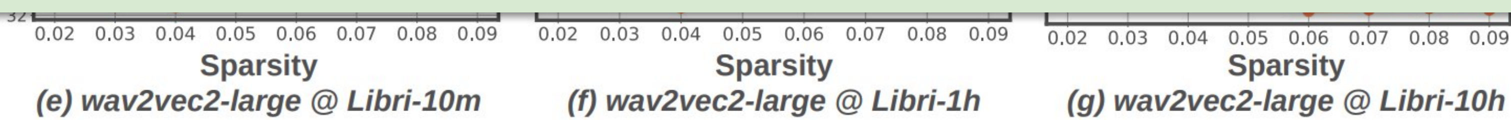

***For example, a 2.34% WER reduction achieved at a 8% sparsity ratio on wav2vec2-base/Libri-10m***

# S³-Router's App. 1: A New Finetuning Paradigm

- **Discarding ≤10% weights is all you need**
  - Consistently outperform the standard weight finetuning in terms of the achievable word error rate (WER)




**Key Insight: Tuning model connections instead of weights can reduce overfitting on low-resource speech**



# $S^3$-Router's App. 1: A New Finetuning Paradigm

- **Discarding ≤10% weights is all you need**
  - Consistent phoneme error rate (PER) reductions over standard weight finetuning for *cross-lingual transfer*
    - *Setup*: Finetune wav2vec2-base on CommonVoice

| Language | Dutch | Mandarin | Spanish | Tatar | Russian |
|---|---|---|---|---|---|
| Weight ft | 19.82 | 26.67 | 13.86 | 11.14 | 17.05 |
| $S^3$-Router | **18.51** | **26.10** | **13.37** | **10.94** | **16.33** |
| **Language** | **Italian** | **Kyrgyz** | **Turkish** | **Swedish** | **France** |
| Weight ft | 19.27 | 13.41 | 15.70 | 20.81 | 19.35 |
| $S^3$-Router | **18.29** | **12.30** | **14.82** | **19.64** | **17.94** |

# $S^3$-Router's App. 1: A New Finetuning Paradigm

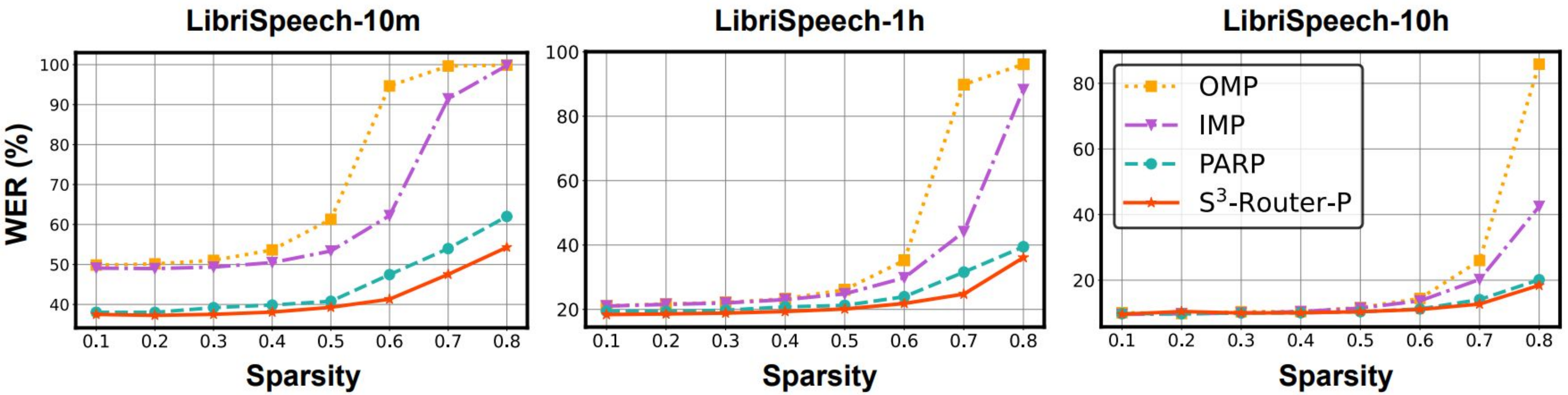
- **Discarding ≤10% weights is all you need**
  - Consistent phoneme error rate (PER) reductions over standard weight finetuning for *cross-lingual transfer*
    - *Setup*: Finetune wav2vec2-base on CommonVoice

| Language | Dutch | Mandarin | Spanish | Tatar | Russian |
|---|---|---|---|---|---|
| Weight ft | 19.27 | 13.41 | 15.70 | 20.81 | 19.35 |
| $S^3$-Router | **18.29** | **12.30** | **14.82** | **19.64** | **17.94** |

***Boost Multilingual Efficiency*: $S^3$-Router can simultaneously support aforementioned 11 languages with -88.5% parameters**

# $S^3$-Router's App. 2: A SOTA Pruning Scheme

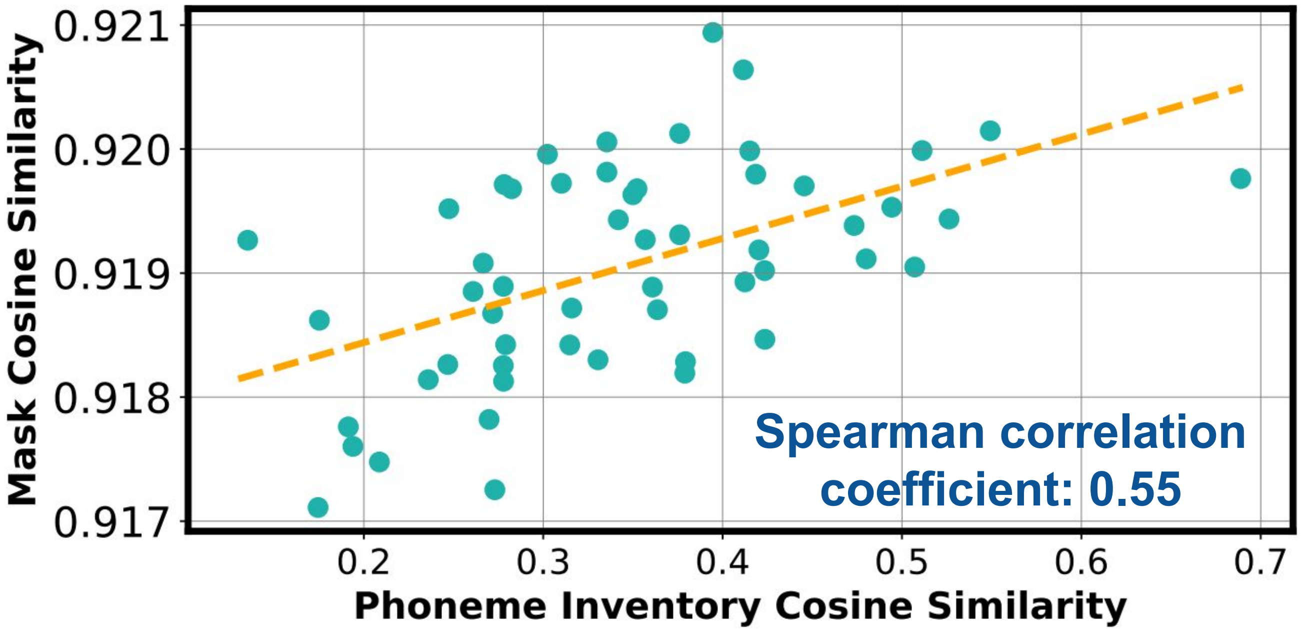
- **Observation:** Achieve better or comparable pruning effectiveness over SOTA ASR pruning techniques




***For example, a 6.46% lower WER over PARP [NeurIPS'21] under a sparsity ratio of 70% with only 10min labeled data***

# $S^3$-Router's App. 3: Analyze Speech SSL Models

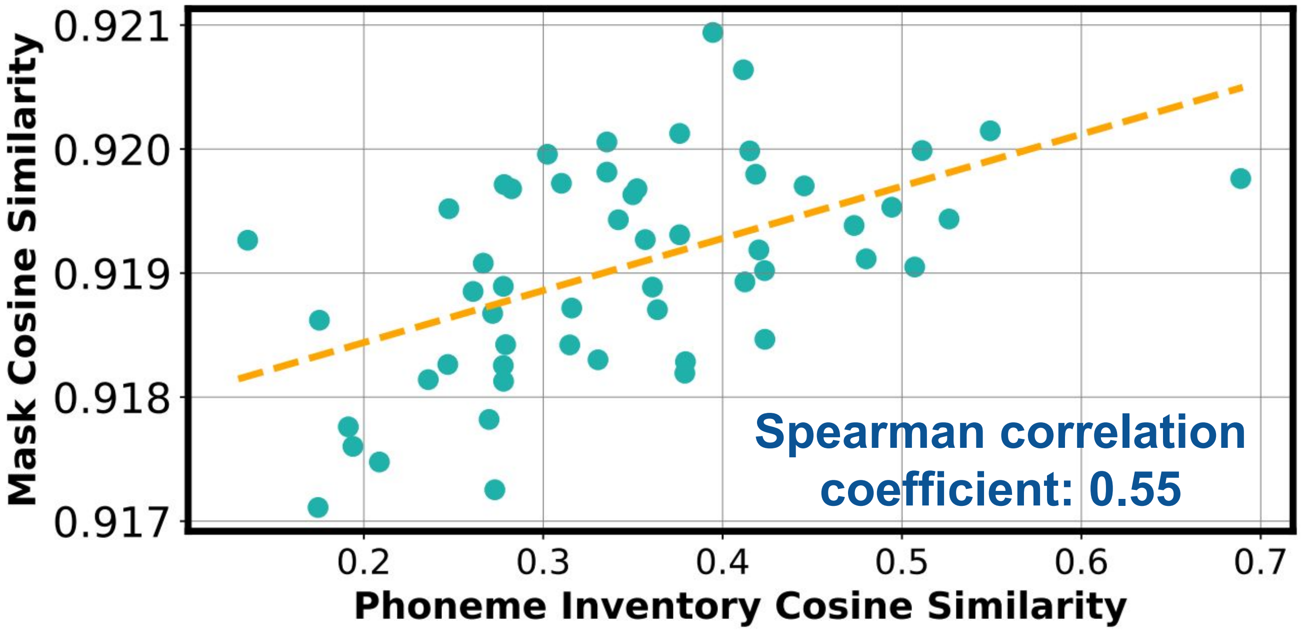
- **How is the learned masks correlated to phonetics?**
  - Visualize the correlation between the mask similarity and the phonetic similarity of different languages




**$S^3$-Router can be utilized to analyze the encoded phonetic differences between languages from speech SSL models' views**

# $S^3$-Router's App. 3: Analyze Speech SSL Models

- **How is the learned masks correlated to phonetics?**
  - Visualize the correlation between the mask similarity and the phonetic similarity of different languages




***Much more experiments* are provided in our paper!**

NEURAL INFORMATION PROCESSING SYSTEMS

EFFICIENT AND INTELLIGENT LAB

# Losses Can Be Blessings: Routing Self-Supervised Speech Representations Towards Efficient Multilingual and Multitask Speech Processing

*NeurIPS 2022*

The work is supported by the National Science Foundation (NSF) through the CCRI program and an IBM faculty award.

Georgia Tech | School of Computer Science
College of Computing

IBM | MIT