Towards Zero-Shot Code-Switched Speech Recognition

Brian Yan, Matthew Wiesner, Ondrej Klejch, Preethi Jyothi, Shinji Watanabe

The Zero-Shot Code-Switching Problem

- Need to generalize: utterance level LID → intra-sentential LID
- Efficiently leverage CS text data, if available

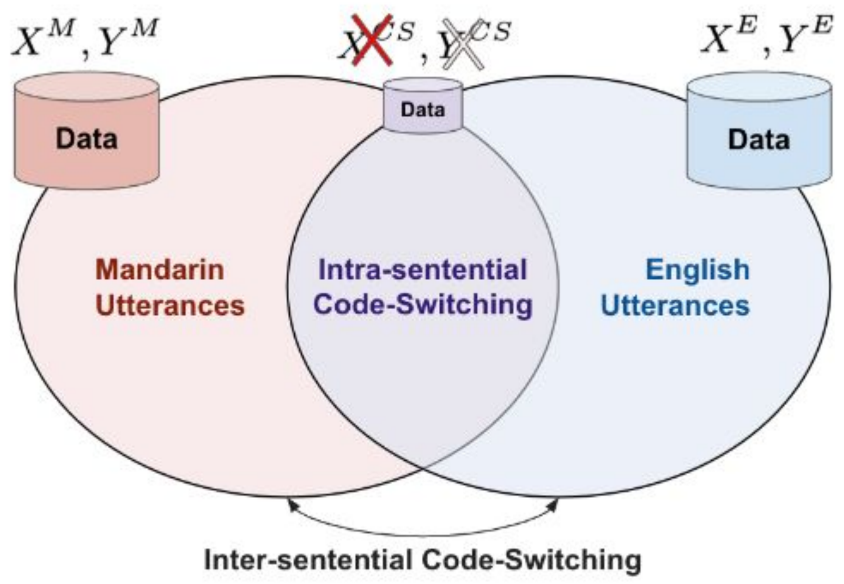
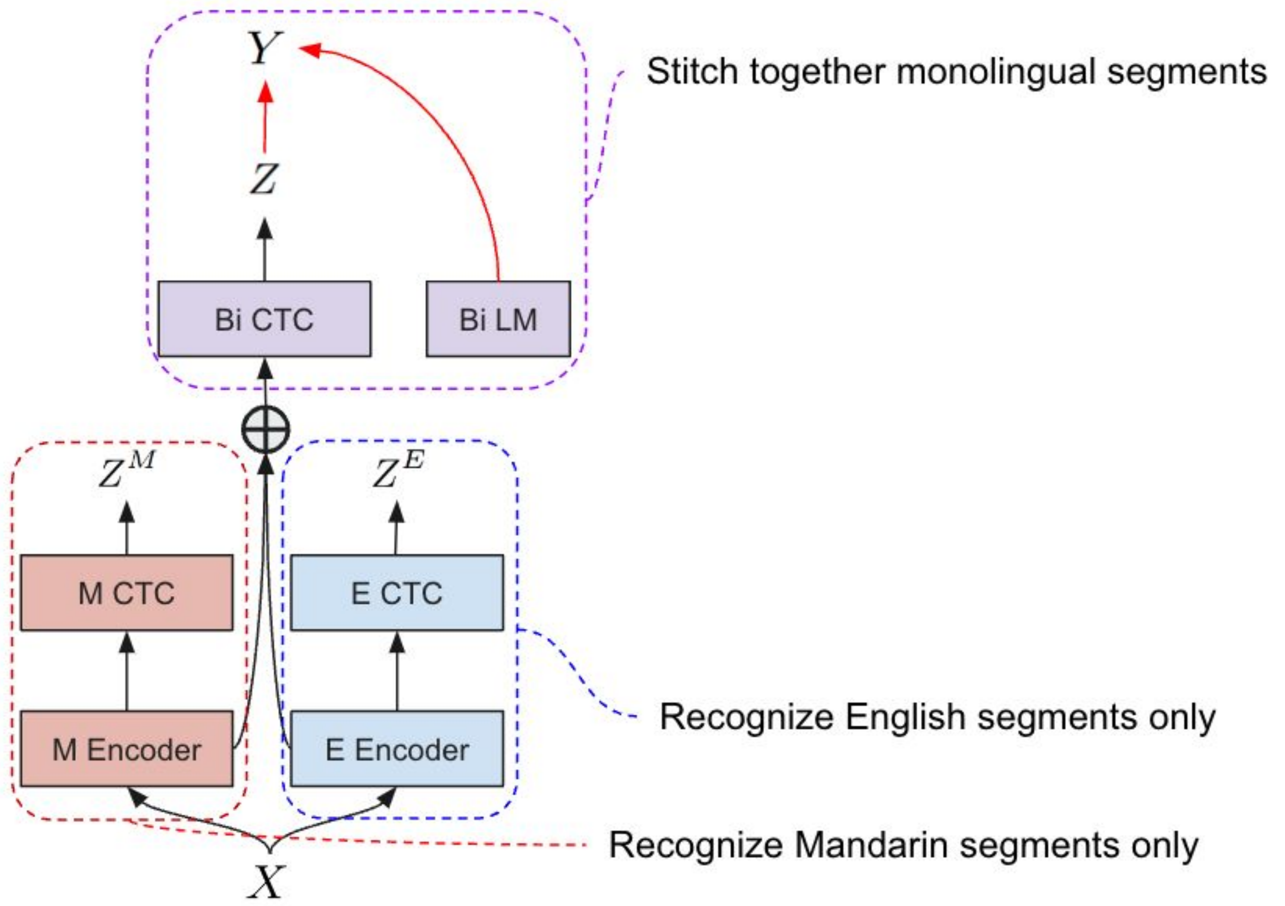


Table 1. SEAME train, devman, and devsge sets broken down by language with hours of duration and number of sentences for speech and text data respectively. †Allowed in fully zero-shot settings. *Original train split [38] was up-sampled by 3x via 0.9 and 1.1 speed perturbations [39].

Set	Type	Full	CS	Mono
TRAIN*	Speech	303h	204h	99h†
TRAIN	Text	89k	50k	39k†
DEVMAN	Speech	8h	6h	2h
DEVSGE	Speech	4h	2h	2h

Conditional Code-Switching Framework

- Conditional approaches are strong in fully supervised settings (prior works)
  - **Monolingual experts**: data efficient, reduced monolingual/CS interference



$$p(Y|X) \approx \underbrace{p(Y)}_{\triangleq p_{\text{Bi\_LM}}(Y)} \underbrace{\sum_{Z} p(Z|Z^M, Z^E)}_{\triangleq p_{\text{Bi\_CTC}}(Y|Z^M, Z^E)} \underbrace{\sum_{Z^M} p(Z^M|X)}_{\triangleq p_{\text{M\_CTC}}(Y^M|X)} \underbrace{\sum_{Z^E} p(Z^E|X)}_{\triangleq p_{\text{E\_CTC}}(Y^E|X)}$$

$Y|X^{CS}$ = _account _ing 还 有

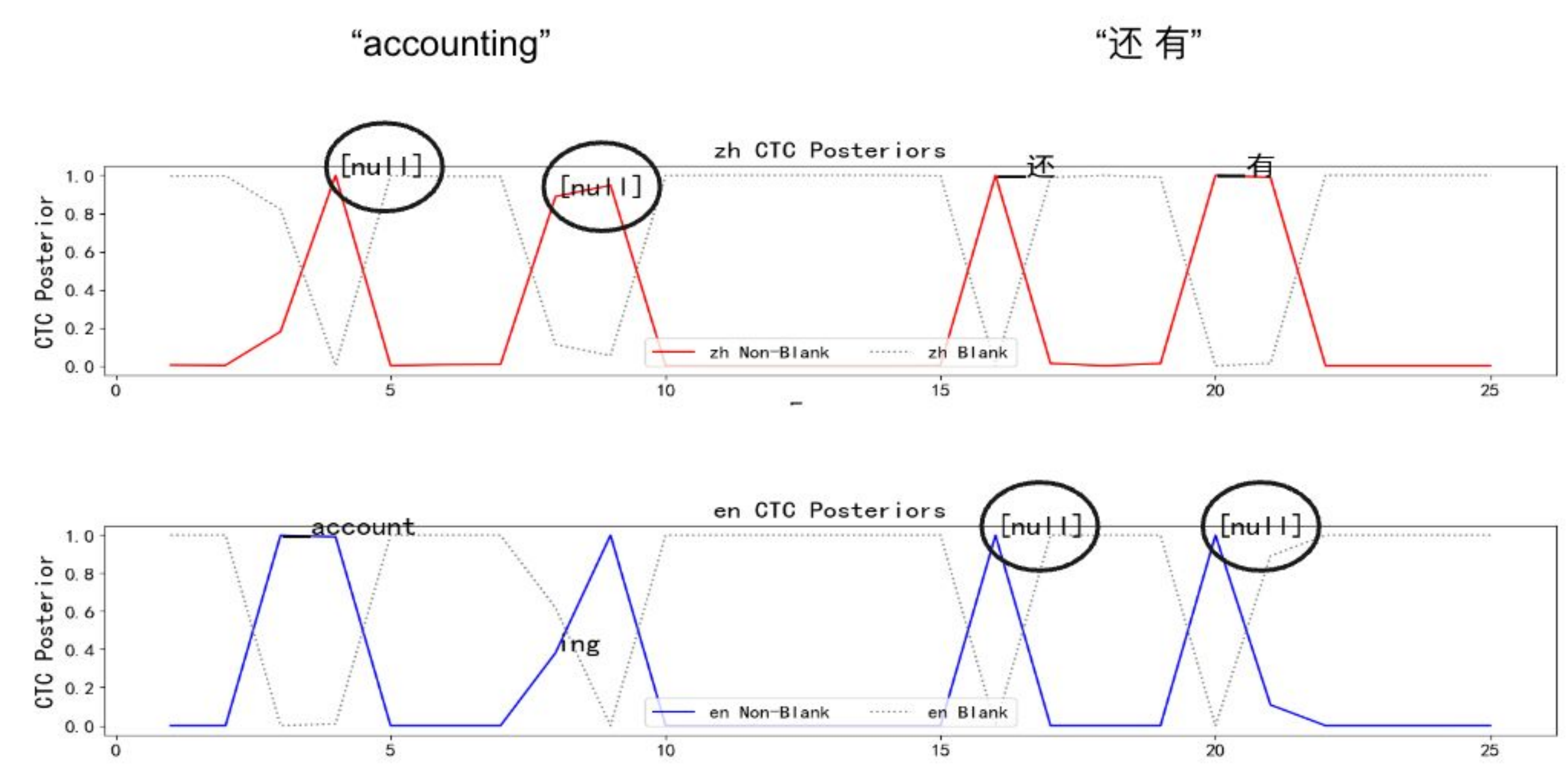
$Y^M|X^{CS}$ = [null] [null] 还 有

$Y^E|X^{CS}$ = _account _ing [null] [null]

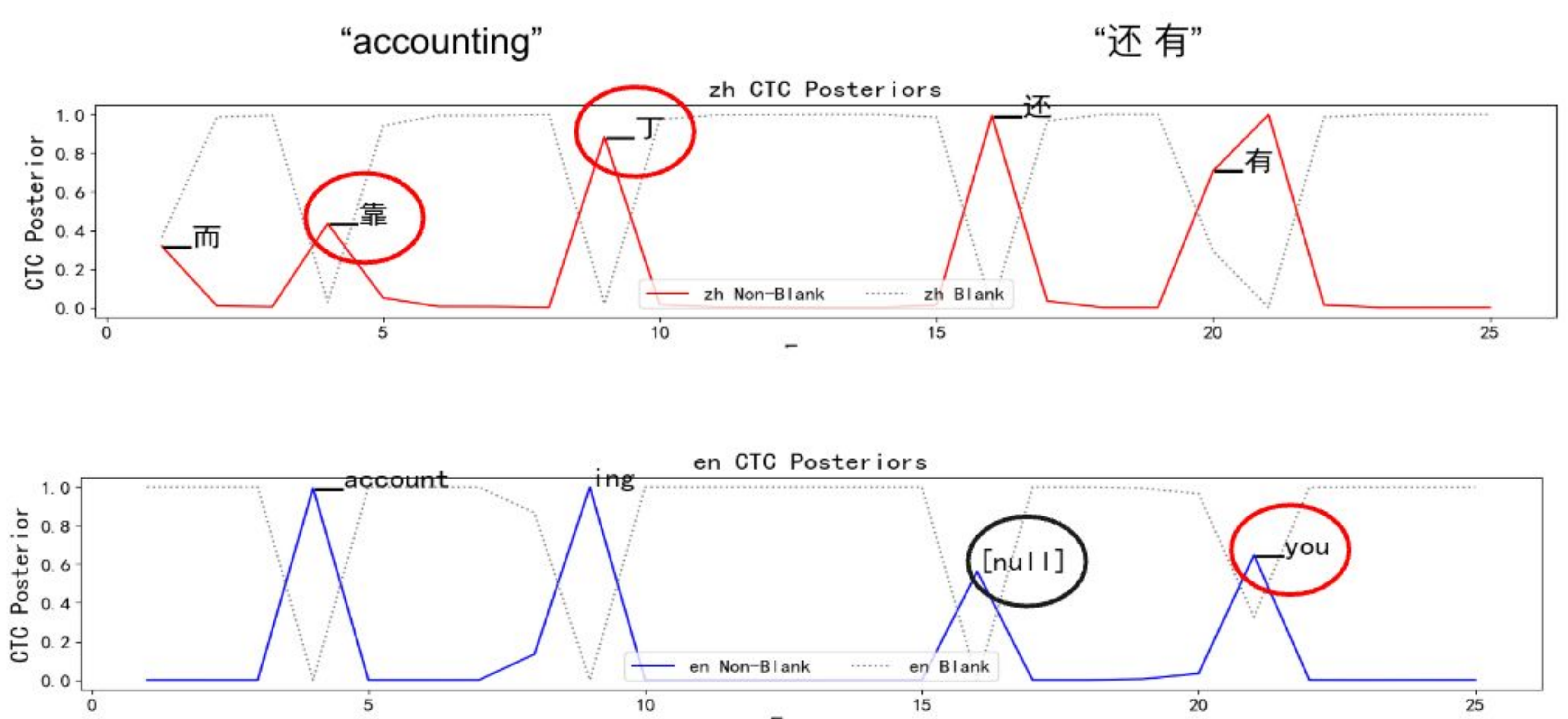
$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{B\_CTC}} + (1-\lambda_1)(\mathcal{L}_{\text{M\_CTC}} + \mathcal{L}_{\text{E\_CTC}})/2.$$

Early Language Segmentation is Fragile

- When trained on **CS** data, monolingual experts can perform **language segmentation**



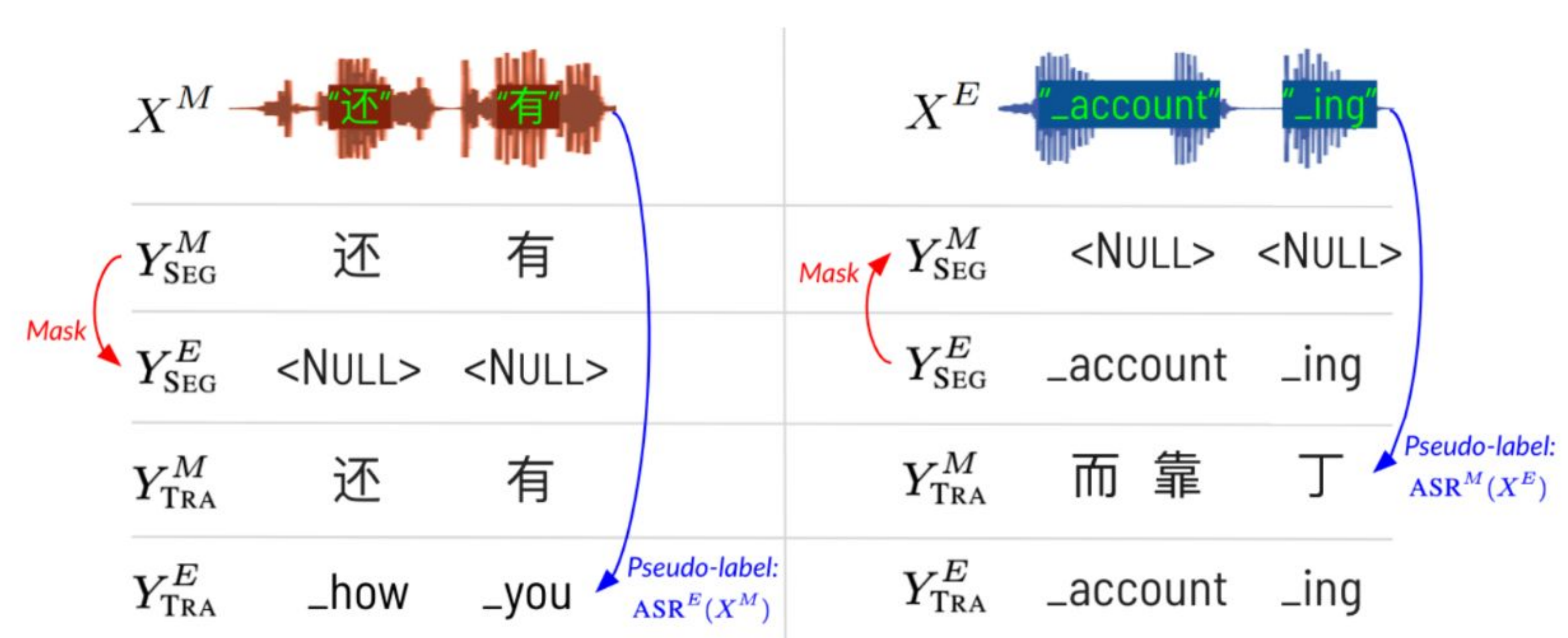
- When trained on **monolingual** data, **language segmentation is unreliable**
  - Each monolingual expert is operating independently
  - Errors are propagated to bilingual modules; ambiguity not in training



Delayed Language Segmentation

Prior: Use **masking** to generate cross-lingual targets

Proposed: Use **pseudo-labeling** to generate cross-lingual targets

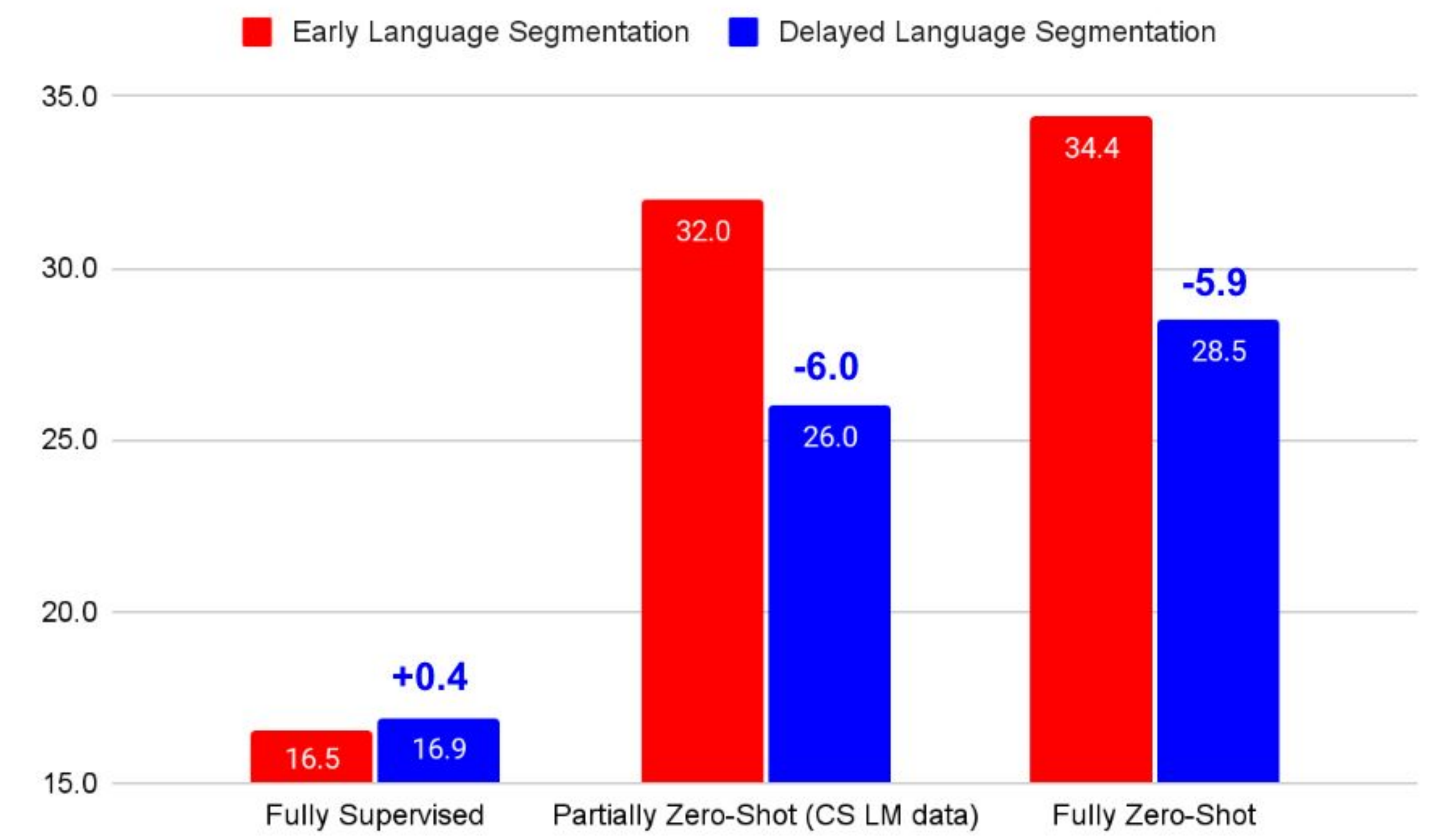


With this change:

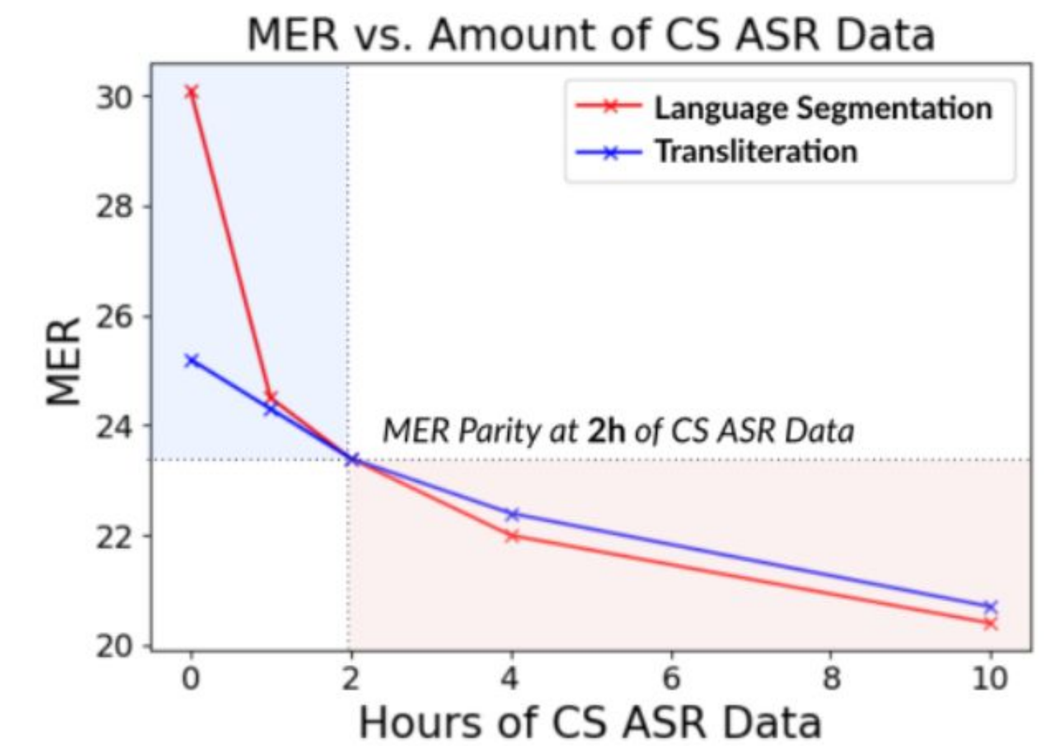
- Monolingual experts *transliterate* the other language (no sense of LID)
- Bilingual modules are responsible for the language segmentation

Zero-Shot Code-Switching: Results

- Delaying language segmentation yields **18% MER** reduction in zero-shot settings
  - Mixed error rate (MER) considers WER for English and CER for Mandarin



- Relaxing the zero-shot setting with CS ASR data, delayed language segmentation is not necessary after **2h** (dataset, language pair dependent)



TLDR

- What did we do?
  - Applied the Conditional CS framework to zero-shot CS ASR, with a simple yet effective training time modification
- General takeaways
  - Language segmentation of code-switched speech is hard, especially if we don't have code-switched supervision
  - Making later decisions about language segmentation is better, allowing us to consider more information (e.g. external LM)

Carnegie Mellon University Language Technologies Institute
Johns Hopkins University
human language technology center of excellence
The University of Edinburgh
Indian Institute of Technology Bombay
byan@cs.cmu.edu