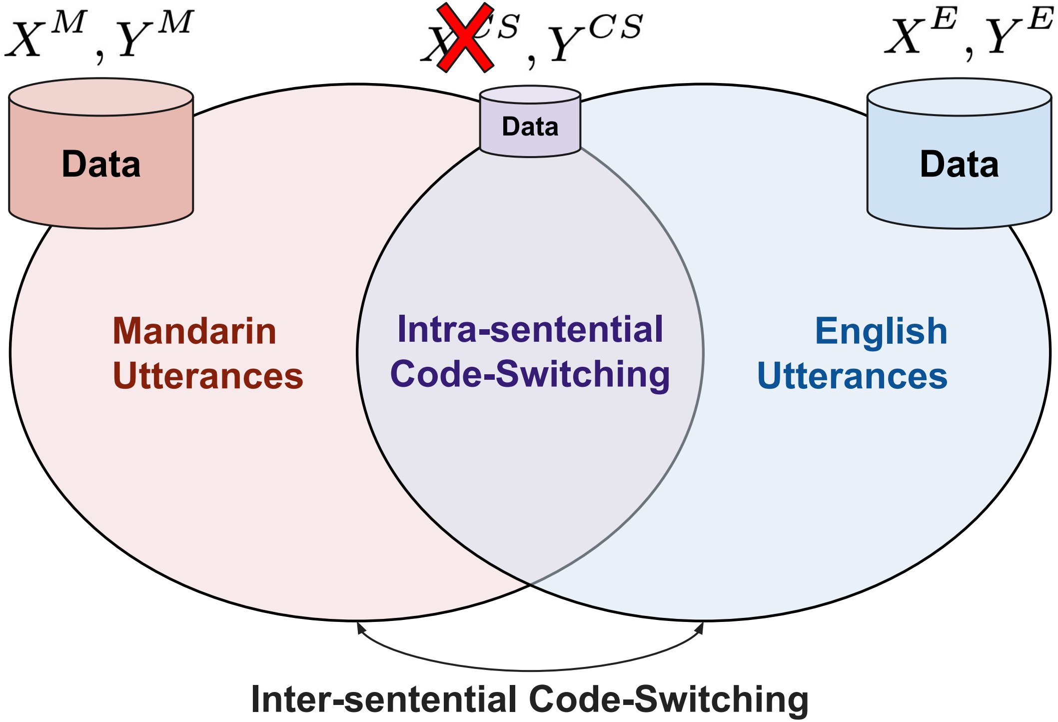
Code-Switched Modeling

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

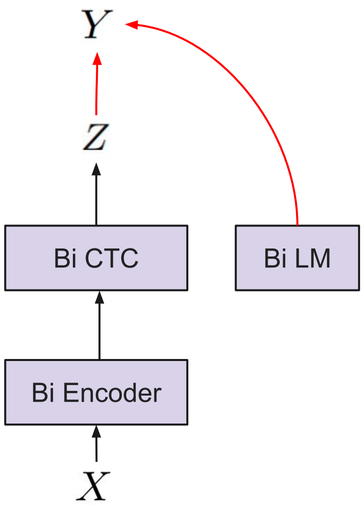
Code-switching (CS) ⊂ Bilingualism

Our objective is to model the **entire bilingual task:**



Joint Modeling of Monolingual and CS ASR

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

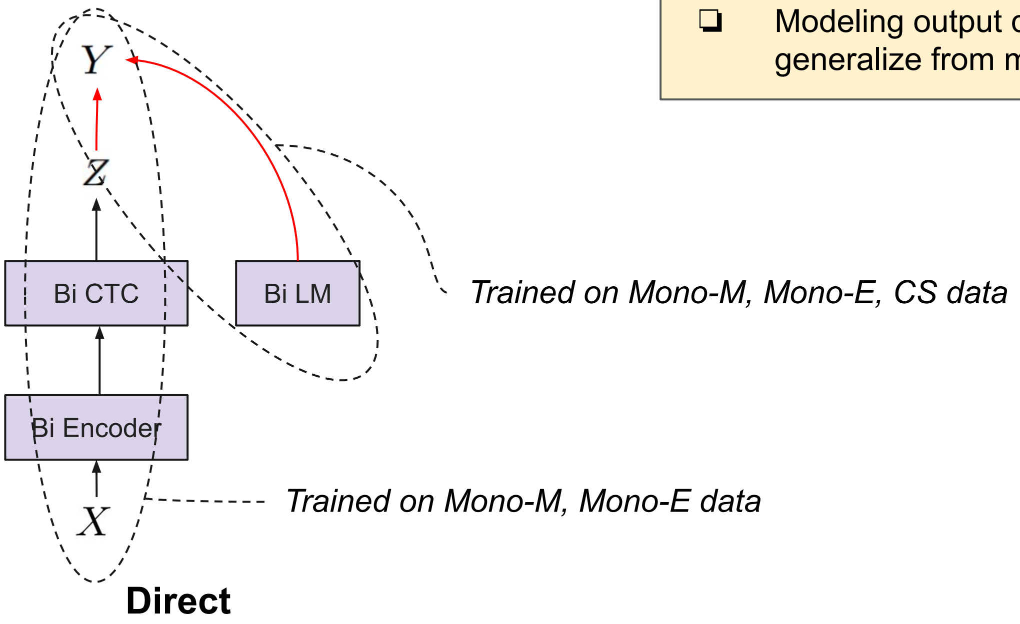


Direct

Bilingual Modules:
handle speech/text which may be Mandarin-only, English-only, or code-switched

Joint Modeling of Monolingual and CS ASR

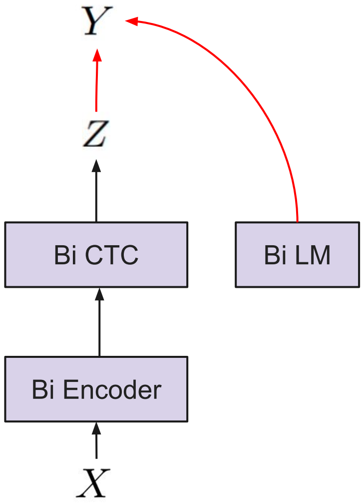
$$p(Y|X) \approx \underbrace{p(Y)}_{\triangleq p_{\text{Bi\_LM}}(Y)} \underbrace{\sum_{\mathcal{Z}} p(Z|X)}_{\triangleq p_{\text{Bi\_CTC}}(Y|X)}$$



- ❏ Modeling output dependency via an external LM → generalize from monolingual ASR training to CS testing

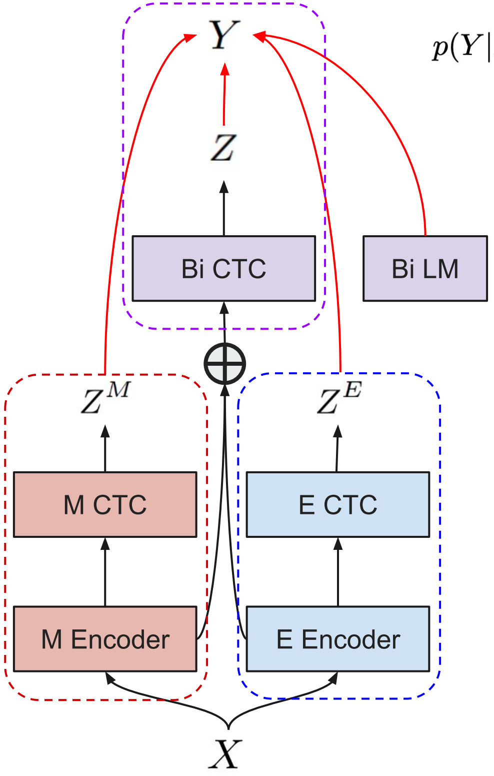
Joint Modeling of Monolingual and CS ASR

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



Direct

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

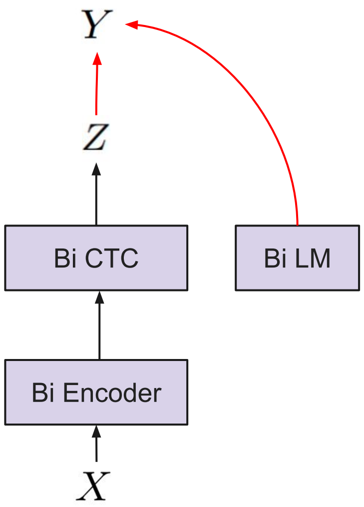


Bilingual Modules +
Monolingual English Expert +
Monolingual Mandarin Expert

Conditionally Factorized

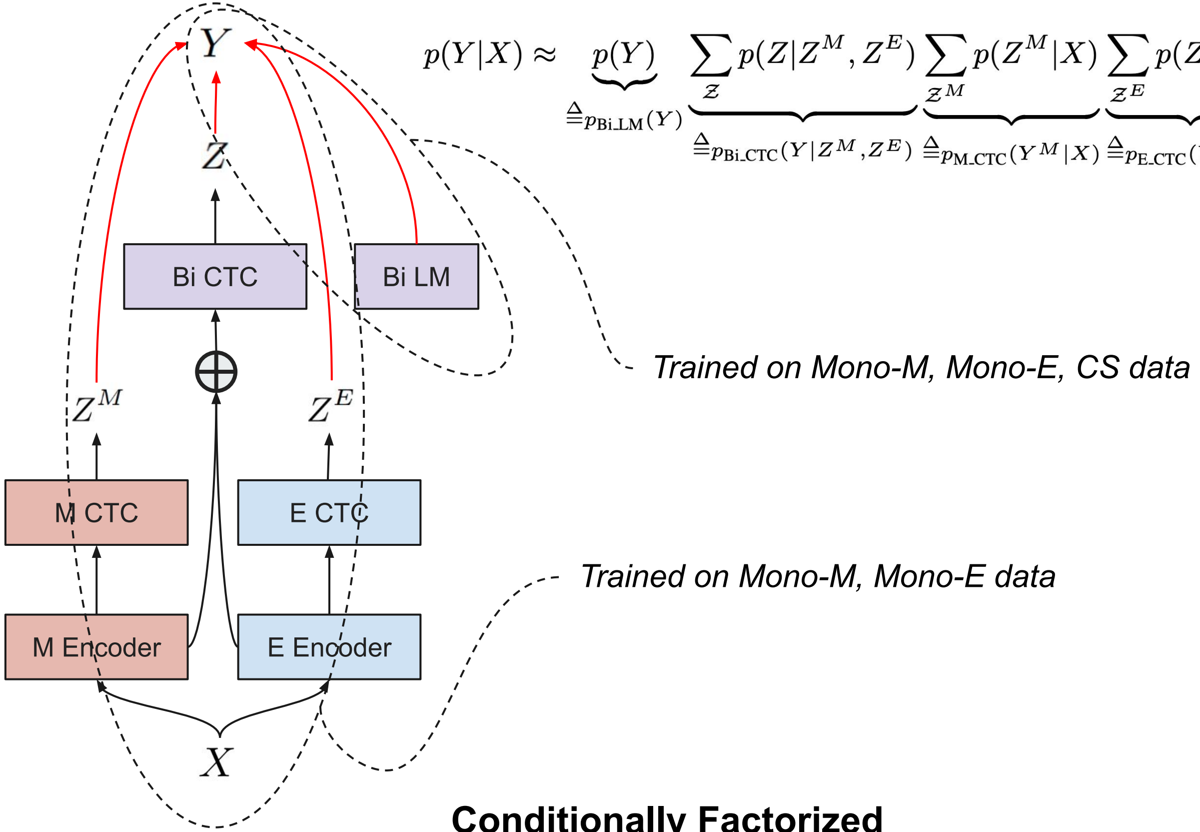
Joint Modeling of Monolingual and CS ASR

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



Direct

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

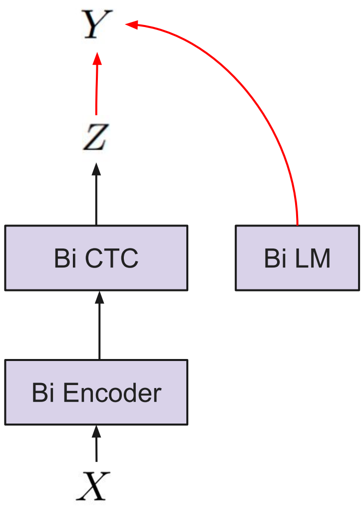


Conditionally Factorized

Brian Yan, Chunlei Zhang, Meng Yu, Shi-Xiong Zhang, Siddharth Dalmia, Dan Berrebbi, Chao Weng, Shinji Watanabe, Dong Yu, “Joint Modeling of Code-Switched and Monolingual ASR via Conditional Factorization”, Proc. ICASSP 2022 (link)

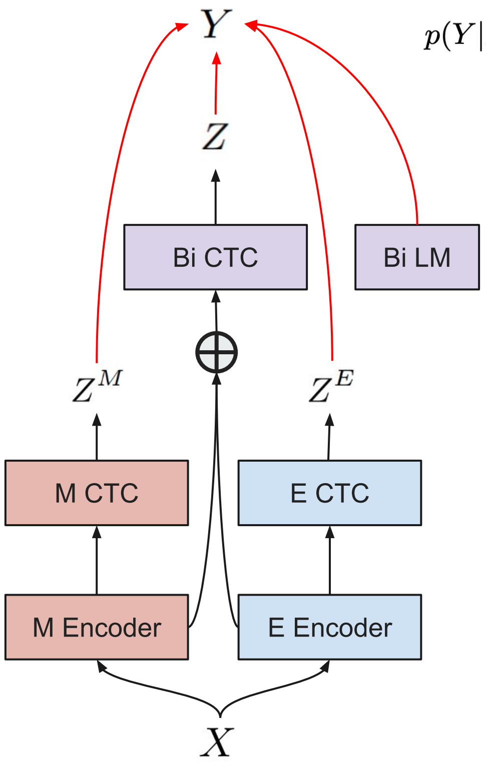
Joint Modeling of Monolingual and CS ASR

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



Direct

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

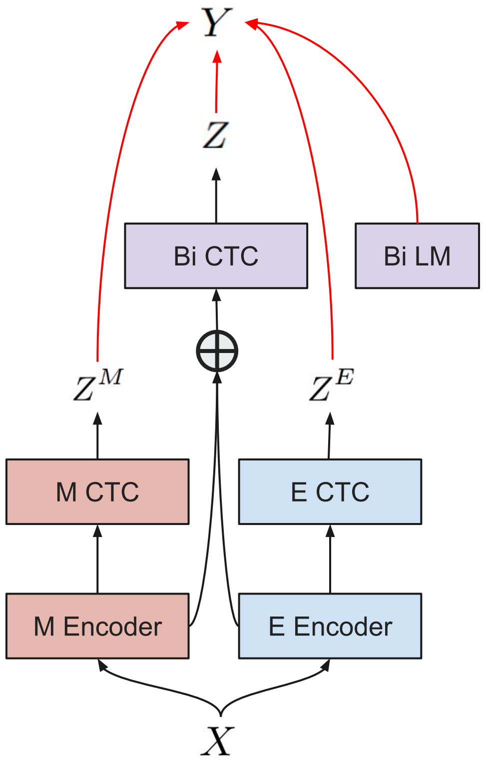


- ❏ Dedicated monolingual sub-components → data efficient training
- ❏ Re-framed the bilingual task → choosing the language per z_i given monolingual information

Conditionally Factorized

Brian Yan, Chunlei Zhang, Meng Yu, Shi-Xiong Zhang, Siddharth Dalmia, Dan Berrebbi, Chao Weng, Shinji Watanabe, Dong Yu, "Joint Modeling of Code-Switched and Monolingual ASR via Conditional Factorization", Proc. ICASSP 2022 (link)

Joint Modeling of Monolingual and CS ASR



Training Scheme

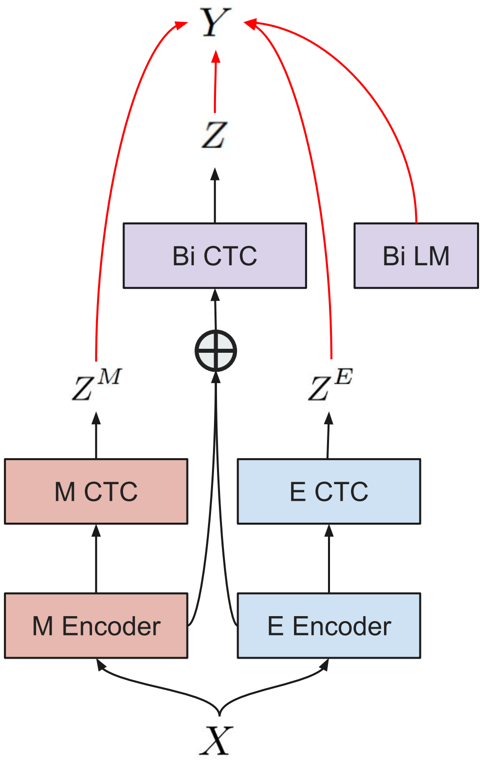
$Y|X^E =$ _account _ing $\qquad Y|X^M =$ 还 有

$Y^M|X^E =$ [null] [null] $\qquad Y^M|X^M =$ 还 有

$Y^E|X^E =$ _account _ing $\qquad Y^E|X^M =$ [null] [null]

$$\mathcal{L}_{\text{LS}} = \lambda\mathcal{L}_{\text{Bi\_CTC}} + (1-\lambda)(\mathcal{L}_{\text{M\_CTC}} + \mathcal{L}_{\text{E\_CTC}})$$

Joint Modeling of Monolingual and CS ASR

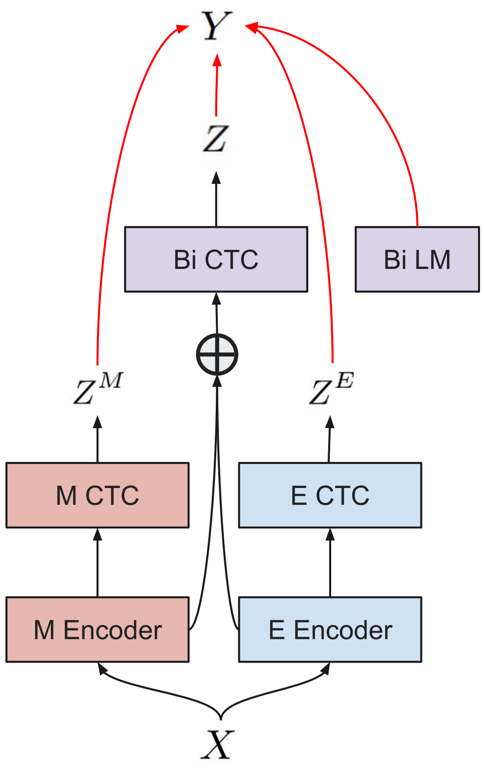


Training Scheme

$Y \mid X^{CS} =$	_account	_ing	还	有
$Y^M \mid X^{CS} =$	[null]	[null]	还	有
$Y^E \mid X^{CS} =$	_account	_ing	[null]	[null]

$$\mathcal{L}_{\text{LS}} = \lambda \mathcal{L}_{\text{Bi\_CTC}} + (1-\lambda)(\mathcal{L}_{\text{M\_CTC}} + \mathcal{L}_{\text{E\_CTC}})$$

Joint Modeling of Monolingual and CS ASR



Training Scheme

$Y \mid X^{CS} =$	_account	_ing	还	有
$Y^M \mid X^{CS} =$	[null]	[null]	还	有
$Y^E \mid X^{CS} =$	_account	_ing	[null]	[null]

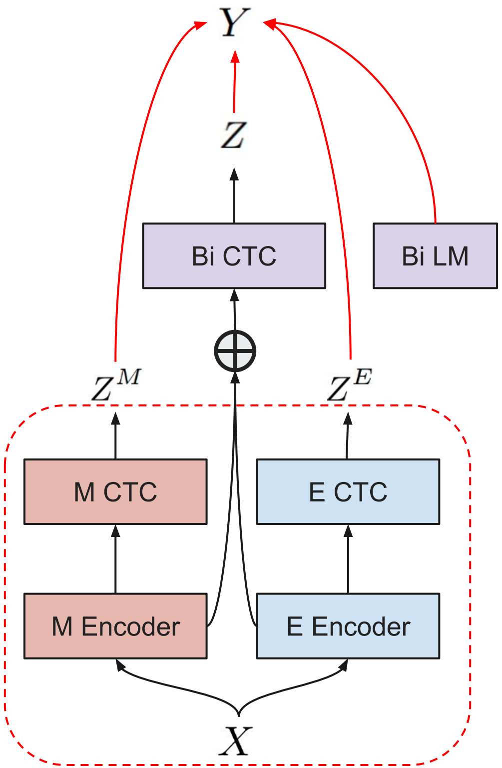
$$\mathcal{L}_{\text{LS}} = \lambda \mathcal{L}_{\text{Bi\_CTC}} + (1-\lambda)(\mathcal{L}_{\text{M\_CTC}} + \mathcal{L}_{\text{E\_CTC}})$$

Inference Procedure

1. Monolingual CTC modules transcribe their respective parts
2. Bilingual CTC module transcribes whole, conditioned on monolingual info.
3. Mono/bilingual CTC modules + bilingual LM jointly decode the final output sequence (e.g. via time sync beam search)

Brian Yan, Chunlei Zhang, Meng Yu, Shi-Xiong Zhang, Siddharth Dalmia, Dan Berrebbi, Chao Weng, Shinji Watanabe, Dong Yu, "Joint Modeling of Code-Switched and Monolingual ASR via Conditional Factorization", Proc. ICASSP 2022 (link)

Joint Modeling of Monolingual and CS ASR



Training Scheme

$Y|X^{CS}$ = _account _ing 还 有

$Y^M|X^{CS}$ = [null] [null] 还 有

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

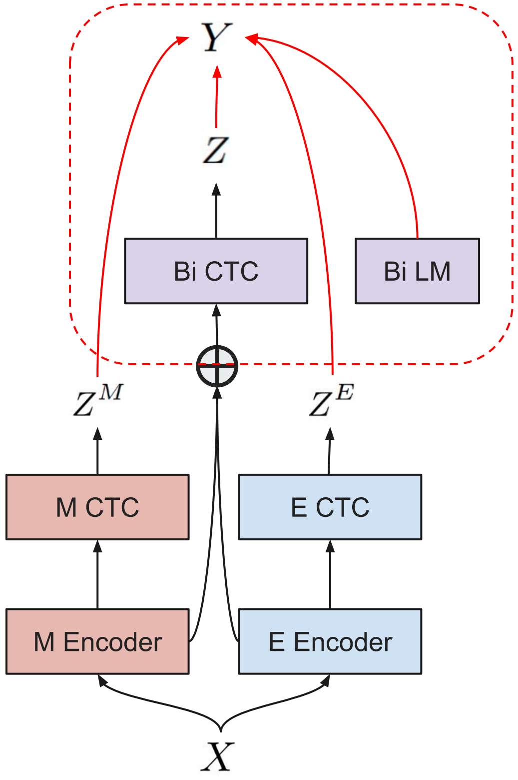
$$\mathcal{L}_{\text{LS}} = \lambda \mathcal{L}_{\text{Bi\_CTC}} + (1-\lambda)(\mathcal{L}_{\text{M\_CTC}} + \mathcal{L}_{\text{E\_CTC}})$$

Inference Procedure

Making a language segmentation decision

1. Monolingual CTC modules transcribe their respective parts
2. Bilingual CTC module transcribes whole, conditioned on monolingual info.
3. Mono/bilingual CTC modules + bilingual LM jointly decode the final output sequence (e.g. via time sync beam search)

Joint Modeling of Monolingual and CS ASR



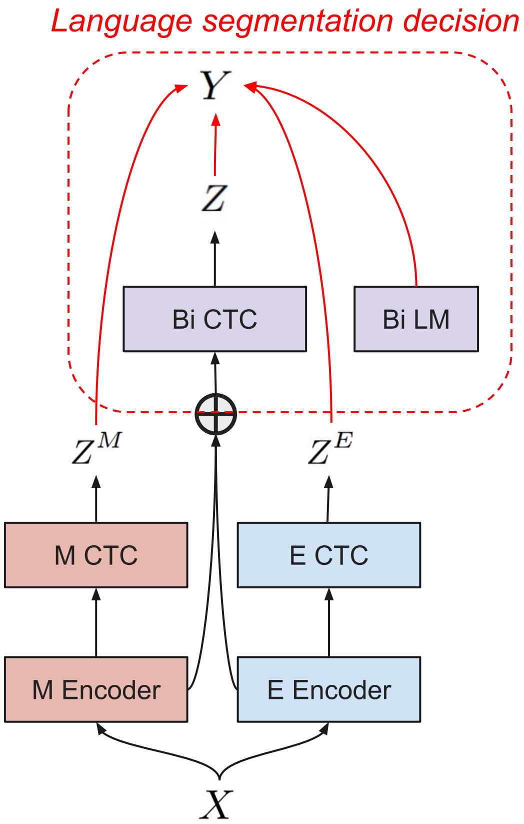
Can we make the language segmentation decision later?

<u>*Inference Procedure*</u>

Making a language segmentation decision

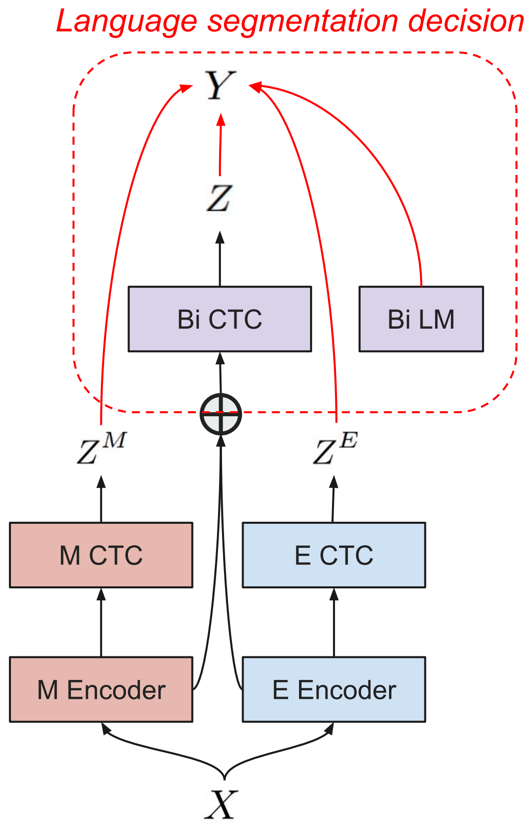
1. Monolingual CTC modules transcribe their respective parts
2. Bilingual CTC module transcribes whole, conditioned on monolingual info.
3. Mono/bilingual CTC modules + bilingual LM jointly decode the final output sequence (e.g. via time sync beam search)

Joint Modeling of Monolingual and CS ASR

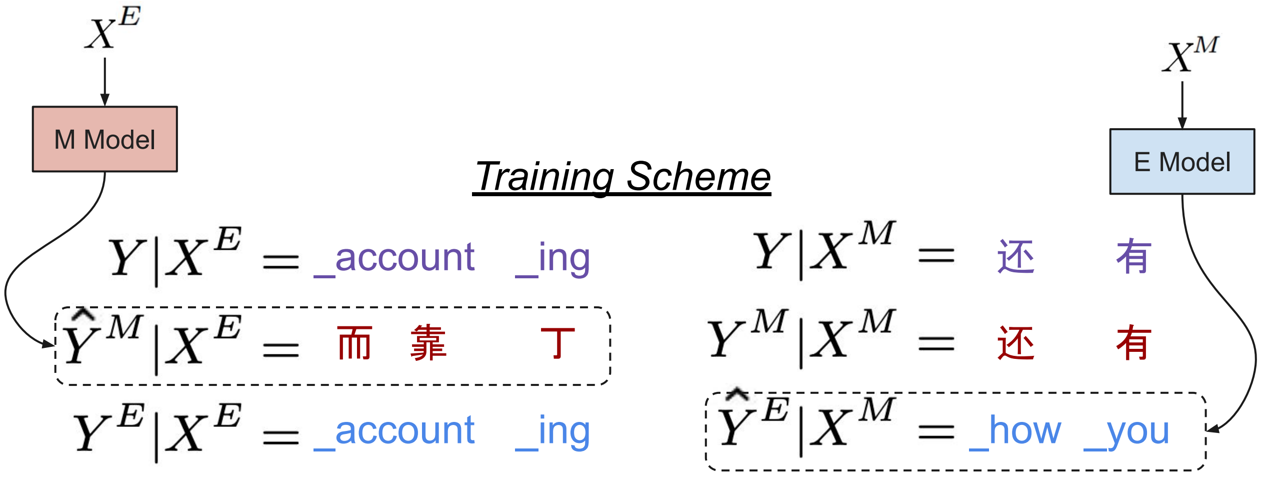


- ❏ Encourage monolingual modules to transcribe the opposite language → leave language segmentation decision to bilingual modules (CTC, LM)

Joint Modeling of Monolingual and CS ASR



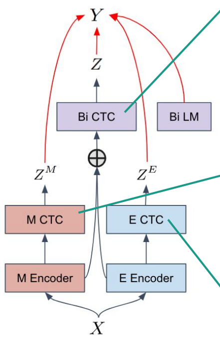
- Encourage monolingual modules to transcribe the opposite language → leave language segmentation decision to bilingual modules (CTC, LM)

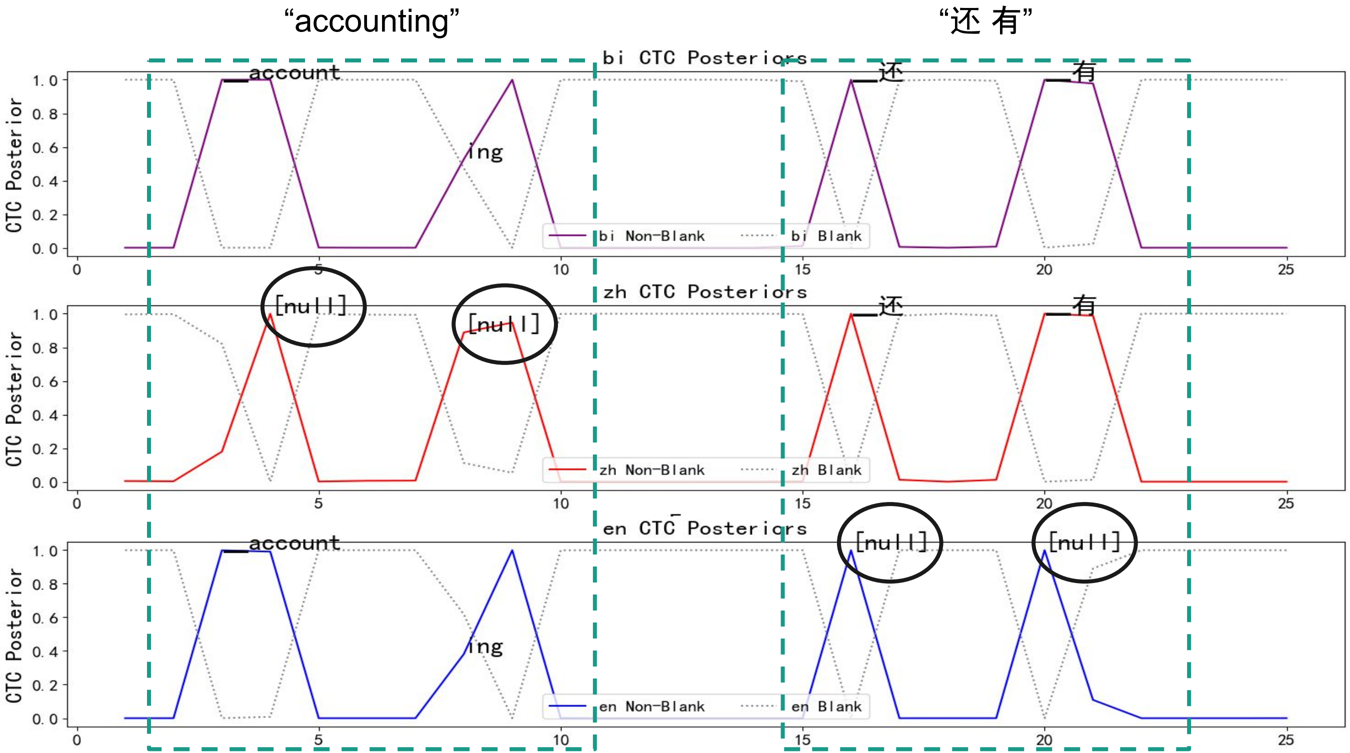


$$\mathcal{L}_{\mathrm{LS}} = \lambda\mathcal{L}_{\mathrm{Bi\_CTC}} + (1-\lambda)(\mathcal{L}_{\mathrm{M\_CTC}} + \mathcal{L}_{\mathrm{E\_CTC}})$$

Qualitative Example: Conditional CTC Posteriors

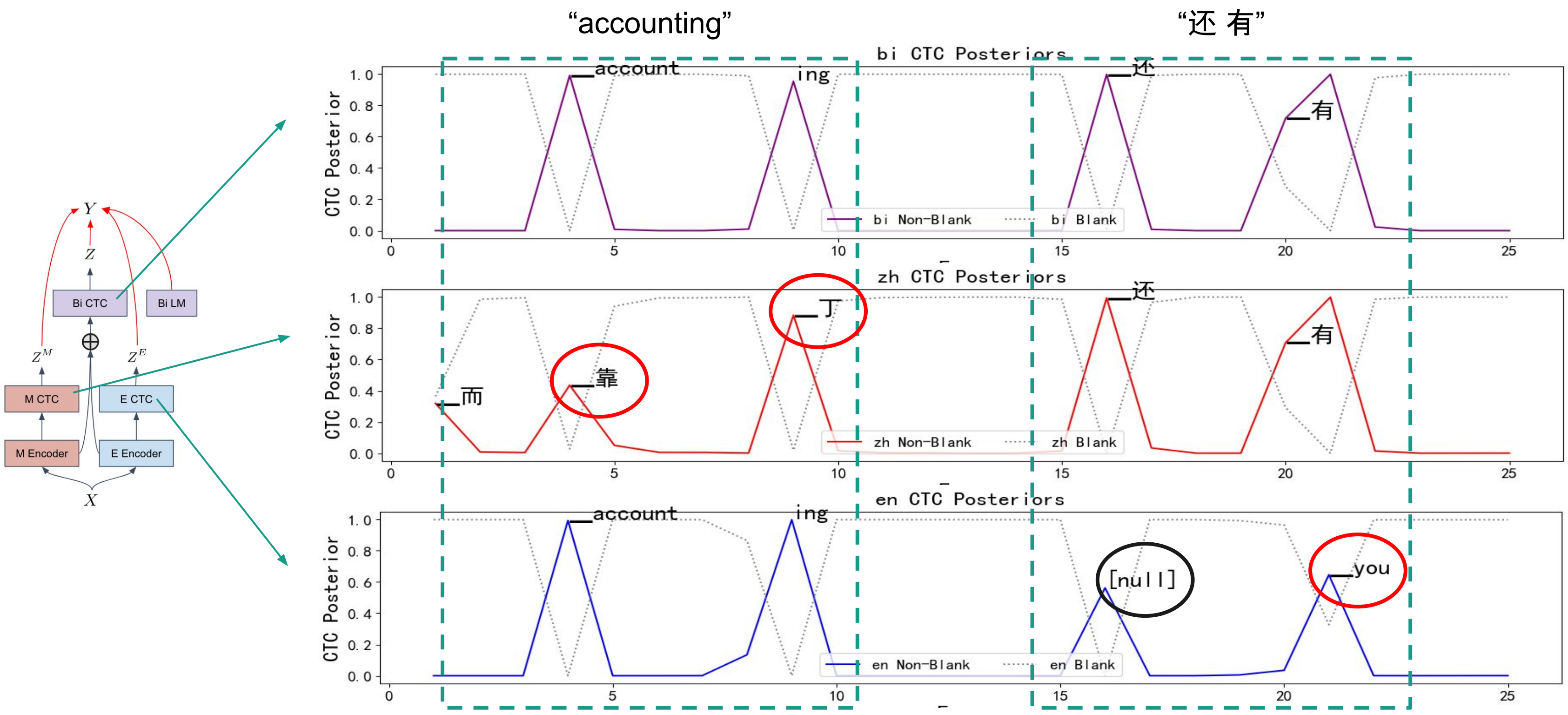
Given CS ASR training data, early language segmentation works well





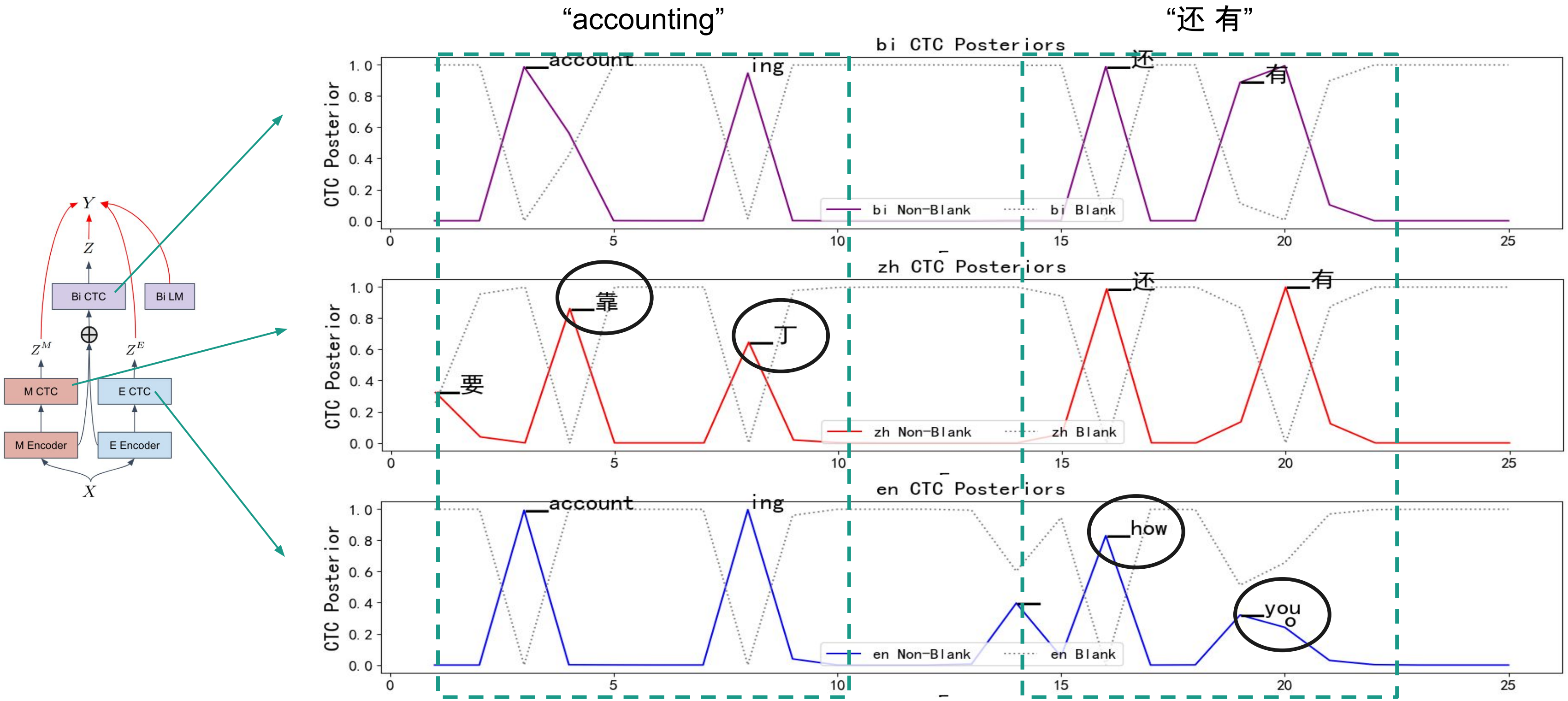
Qualitative Example: Conditional CTC Posteriors

Without CS ASR training data, early language segmentation is unreliable



Qualitative Example: Conditional CTC Posteriors

- Monolingual modules produce smooth likelihoods for opposite lang. instead of [null]
- Language separation information is soft; LM can help decide → late decision



Results

Model	Language Segmentation	ASR Data	LM Data	devman MER(↓)
Conditional CTC	Early	CS + M	-	17.5
Conditional CTC + LM	Early	CS + M	CS + M	**16.8**
Conditional CTC	Early	CS	-	32.3
Conditional CTC + LM	Early	CS	CS + M	30.1
Conditional CTC	Late	CS	-	27.9
Conditional CTC + LM	Late	CS	CS + M	**25.2**

+13.3 MER

-4.9 MER

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)