JOINT MODELING OF CODE-SWITCHED AND MONOLINGUAL ASR VIA CONDITIONAL FACTORIZATION

Brian Yan[1], *Chunlei Zhang*[2], *Meng Yu*[2], *Shi-Xiong Zhang*[2], *Siddharth Dalmia*[1], *Dan Berrebbi*[1], *Chao Weng*[3], *Shinji Watanabe*[1], *Dong Yu*[2]

[1]Carnegie Mellon University, USA, [2]Tencent AI Lab, USA, [3]Tencent AI Lab, China

Carnegie Mellon University Language Technologies Institute

Tencent AI Lab

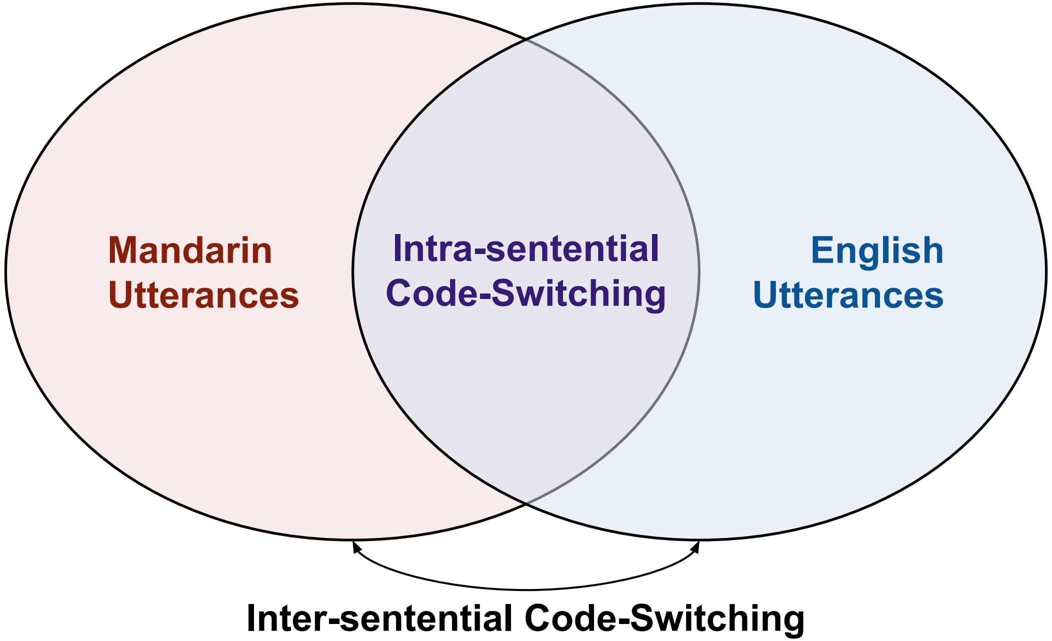
icassp 2022 Singapore China Virtual

Session: SPE-14: Multi-lingual ASR
Session Time: Sunday, 8 May, 23:00 - 23:45 (Singapore Time, UTC +8)

Code-switching (CS) ⊂ Bilingualism

Intra-sentential CS is a **subset** of bilingual conversation, which is often 1 language at a time

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



Bilingual Speech Recognition

Let …

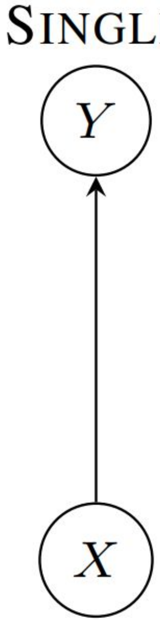
$X = \{\mathbf{x}_t \in \mathbb{R}^D | t = 1, ..., T\}$ denote speech features and

$Y = \{y_t \in (\mathcal{V}^M \cup \mathcal{V}^E) | n = 1, ..., L\}$ denote bilingual transcriptions.

Note that Y may be purely monolingual or code-switched.

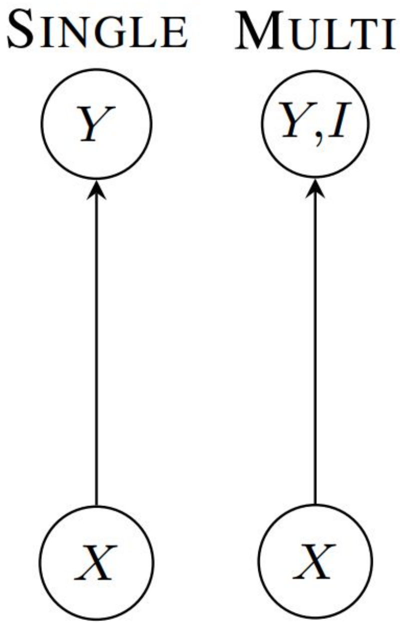
We wish to predict Y given X .

Direct Formulations of Bilingual ASR



- Model Y as a **single** conditionally dependent variable
  - Hybrid: Phone merging (Sivasankaran 2018)
  - E2E: LID token method (Zhang 2020)

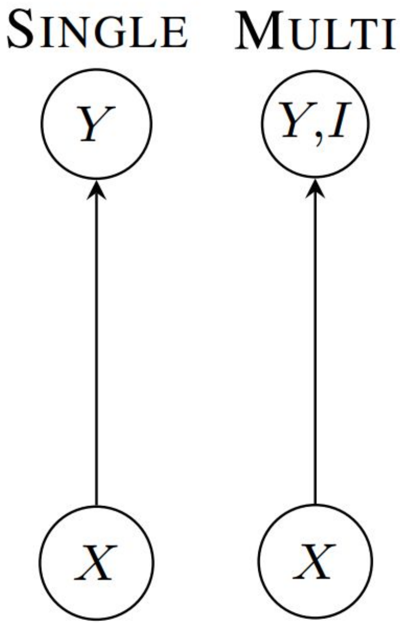
Direct Formulations of Bilingual ASR



- Model *Y* as a **single** conditionally dependent variable
  - Hybrid: Phone merging (Sivasankaran 2018)
  - E2E: LID token method (Zhang 2020)

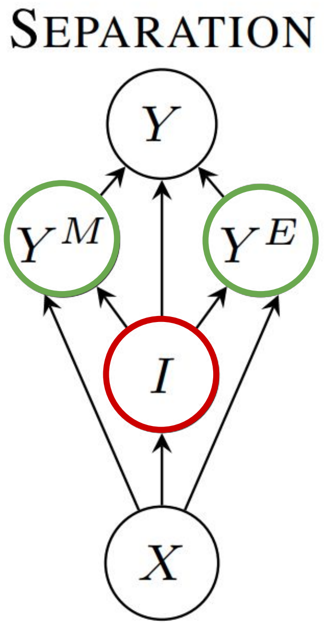
- Model **multiple** dependents: *Y* and language ID, *I*
  - E2E: joint LID and ASR (Zeng 2019)

Direct Formulations of Bilingual ASR



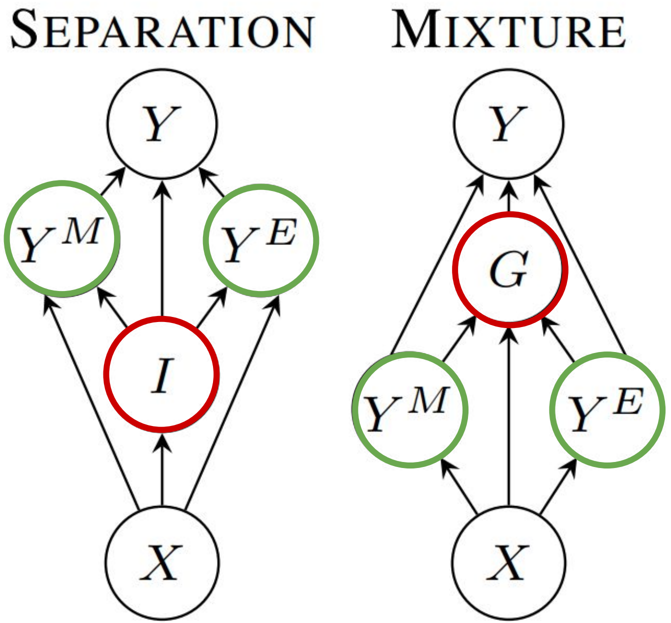
- Model *Y* as a **single** conditionally dependent variable
  - Hybrid: Phone merging (Sivasankaran 2018)
  - E2E: LID token method (Zhang 2020)
- Model **multiple** dependents: *Y* and language ID, *I*
  - E2E: joint LID and ASR (Zeng 2019)
- Combining 2 unrelated languages = more complex

Divide-and-Conquer Formulations of Bilingual ASR



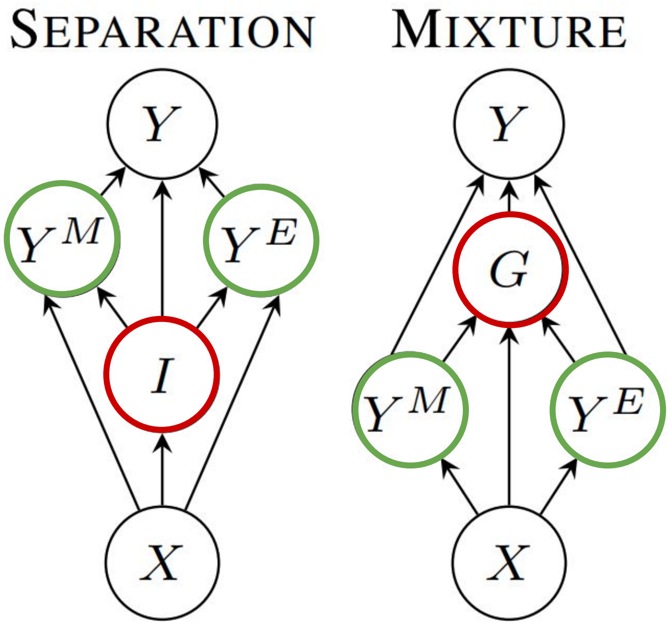
- **Separate** then Recognize
  - Hybrid: LID to monolingual ASR cascade (Chan 2004)

Divide-and-Conquer Formulations of Bilingual ASR



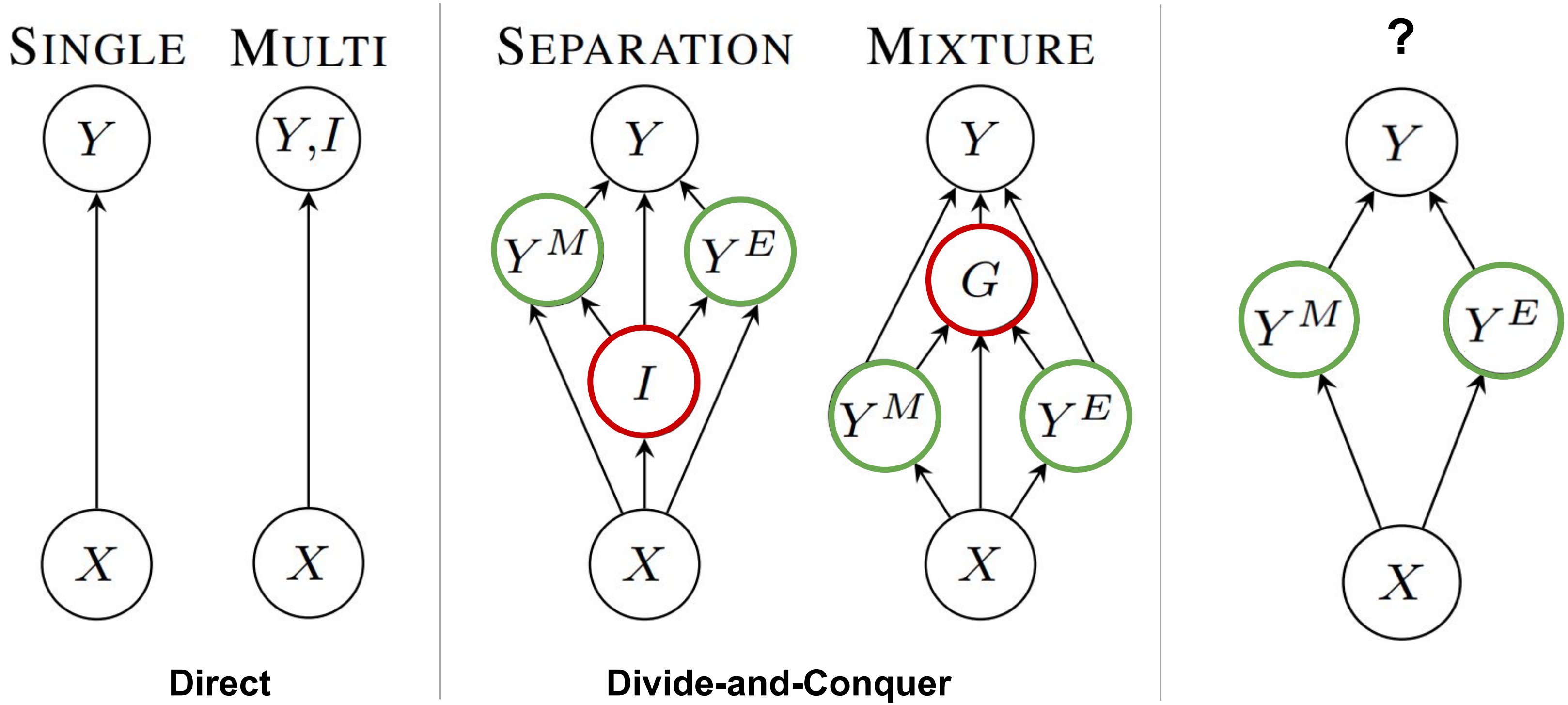
- **Separate** then Recognize
  - Hybrid: LID to monolingual ASR cascade (Chan 2004)
- **Mixture** of Monolingual Experts
  - Hybrid: Frame-level posterior weighting (Weiner 2012)
  - E2E: Mixture of experts (Lu 2020, Dalmia 2021)

Divide-and-Conquer Formulations of Bilingual ASR

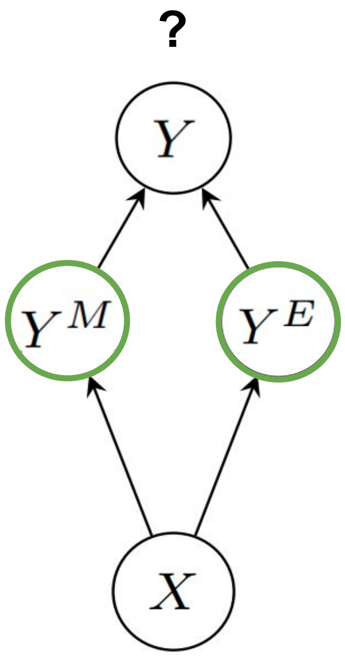


- **Separate** then Recognize
  - Hybrid: LID to monolingual ASR cascade (Chan 2004)
- **Mixture** of Monolingual Experts
  - Hybrid: Frame-level posterior weighting (Weiner 2012)
  - E2E: Mixture of experts (Lu 2020, Dalmia 2021)
- Division of monolingual tasks
  - simpler, more compatible with monolingual data
- Dependence on quality of “divider” module
  - Risk of error propagation, increased complexity

Our Motivation

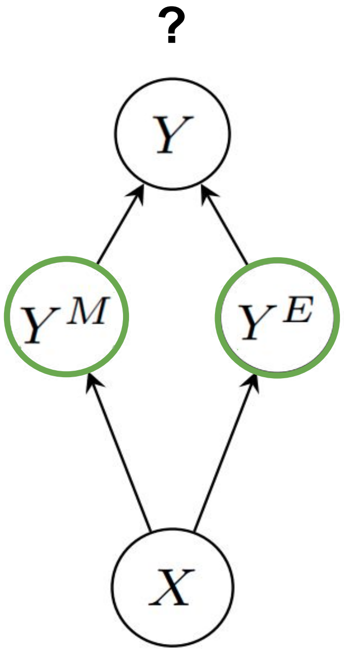


Desiderata



1. Can we build CS + bilingual ASR with **monolingual sub-components**…
2. …where the final output is **conditioned only on those 2 sub-components** and nothing else?
3. And does such a conditional approach more **efficiently leverage monolingual and CS training data**?

Label-to-Frame Synchronization



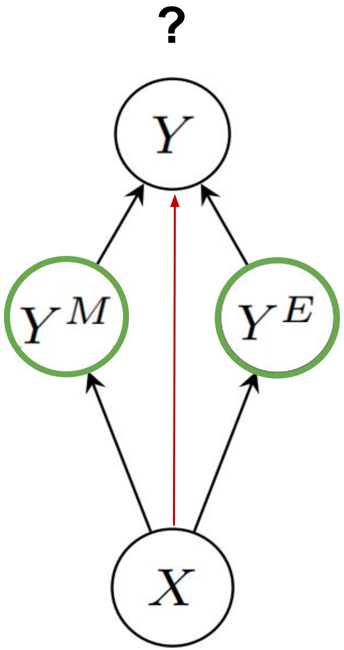
$$P(\mathbf{y}, \mathbf{y}^M, \mathbf{y}^E | \mathbf{x}) = P(\mathbf{y} | \mathbf{y}^M, \mathbf{y}^E) P(\mathbf{y}^M | \mathbf{x}) P(\mathbf{y}^E | \mathbf{x})$$

If …

Y^M = 什么是, Y^E = code-switching

Can Y be determined?

The Need for Label-to-Frame Synchronization



$$P(\mathbf{y},\mathbf{y}^M,\mathbf{y}^E|\mathbf{x}) = \boxed{P(\mathbf{y}|\mathbf{y}^M,\mathbf{y}^E)}P(\mathbf{y}^M|\mathbf{x})P(\mathbf{y}^E|\mathbf{x})$$

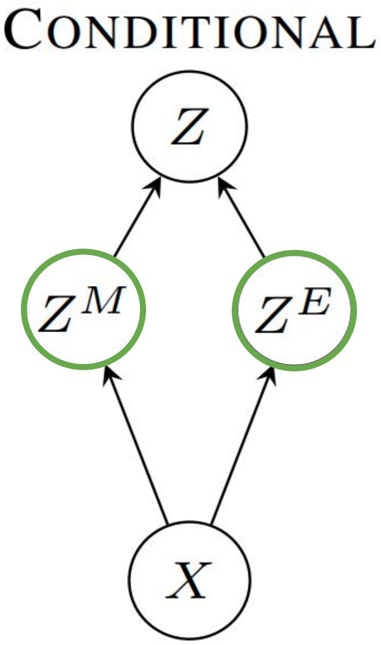
If …

Y^M = 什么是, Y^E = code-switching

Can Y be determined?

→ No! We’re missing ordering information
→ This formulation is not conditionally independent

Conditionally Factorized Bilingual ASR



Let …

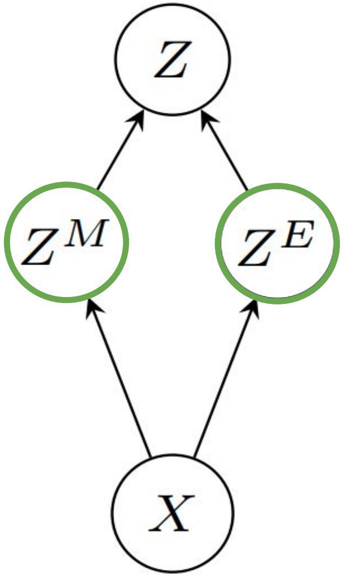
$X = \{\mathbf{x}_t \in \mathbb{R}^D | t = 1, ..., T\}$ denote speech features and

$Z = \{z_t \in \mathcal{V}^M \cup \mathcal{V}^E \cup \{\varnothing\} | t = 1 \ldots T\}$ denote label-to-frame alignments.

Denotes null emision as in CTC

Conditionally Factorized Bilingual ASR

CONDITIONAL



Let …

$X = \{\mathbf{x}_t \in \mathbb{R}^D | t = 1, ..., T\}$ denote speech features and

$Z = \{z_t \in \mathcal{V}^M \cup \mathcal{V}^E \cup \{\varnothing\} | t = 1 \ldots T\}$ denote label-to-frame alignments.

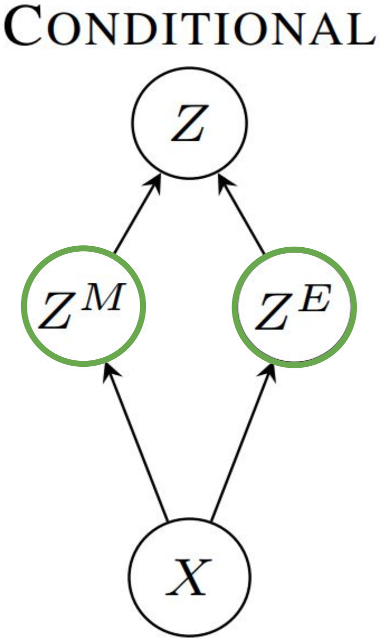
Denotes null emision as in CTC

Bilingual label-to-frame alignments can be specified in terms of its constituent monolingual label-to-frame alignments.

$$z_t = \begin{cases} z_t^M, & \text{if } z_t^M \in \mathcal{V}^M \text{ and } z_t^E = \varnothing \\ z_t^E, & \text{if } z_t^E \in \mathcal{V}^E \text{ and } z_t^M = \varnothing \\ \varnothing, & \text{if } z_t^M = \varnothing \text{ and } z_t^E = \varnothing \end{cases}$$

By definition, at least one side is null for a given t

Conditionally Factorized Bilingual ASR

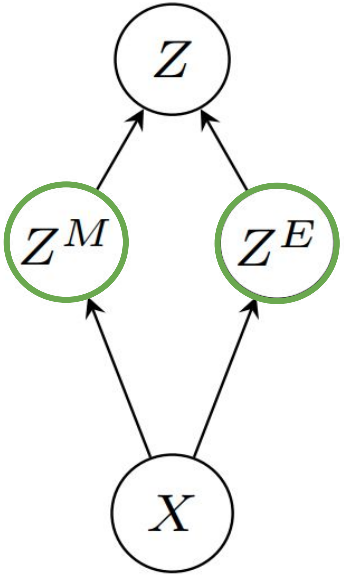


Formulate likelihood in terms of label-to-frame alignments:

$$p(Y|X) = \sum_{Z \in \mathcal{Z}(Y)} p(Z|X)$$

Conditionally Factorized Bilingual ASR

CONDITIONAL



Formulate likelihood in terms of label-to-frame alignments:

$$p(Y|X) = \sum_{Z \in \mathcal{Z}(Y)} p(Z|X)$$

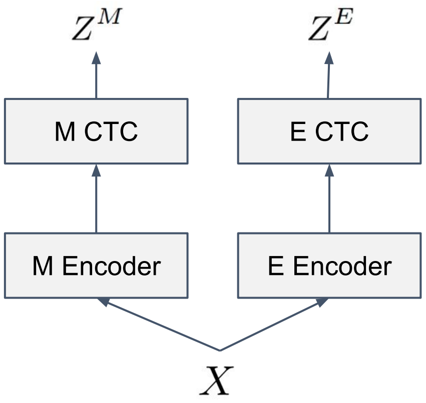
Jointly model CS and Monolingual parts, w/ **conditional factorization**:

$$\begin{aligned} p(Z|X) &= p(Z, Z^M, Z^E|X) \\ &= p(Z|Z^M, Z^E, X)p(Z^M, Z^E|X) \\ &\approx p(Z|Z^M, Z^E, \cancel{X})p(Z^M|X)p(Z^E|X) \end{aligned}$$

Independence

Conditional independence

Conditional RNN Transducer

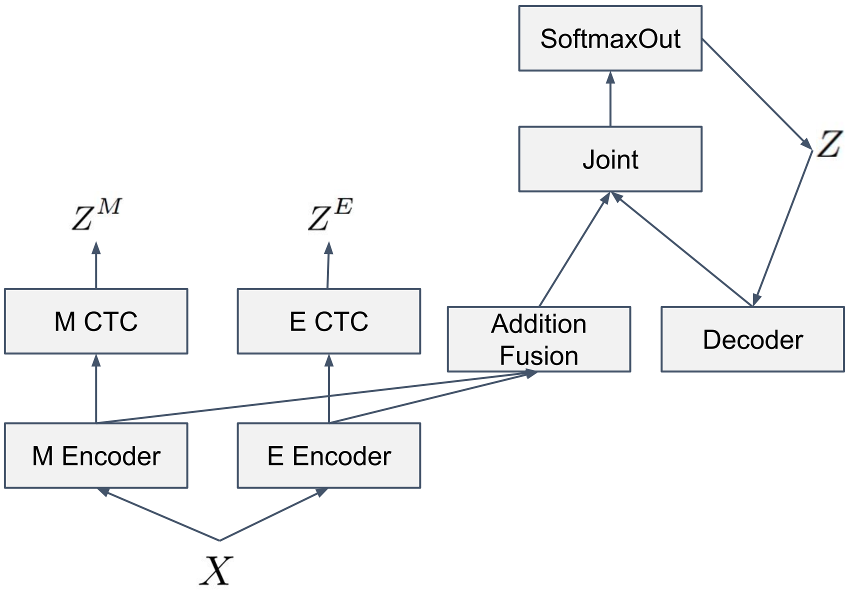


Monolingual CTC Modules:

$$p(Z^M|X) \approx \prod_{t=1}^{T} p(z_t^M|X, \cancel{z_{1:t-1}^M})$$

$$p(Z^E|X) \approx \prod_{t=1}^{T} p(z_t^E|X, \cancel{z_{1:t-1}^E})$$

Conditional RNN Transducer



Monolingual CTC Modules:

$$p(Z^M|X) \approx \prod_{t=1}^{T} p(z_t^M|X, \cancel{z_{1:t-1}^M})$$

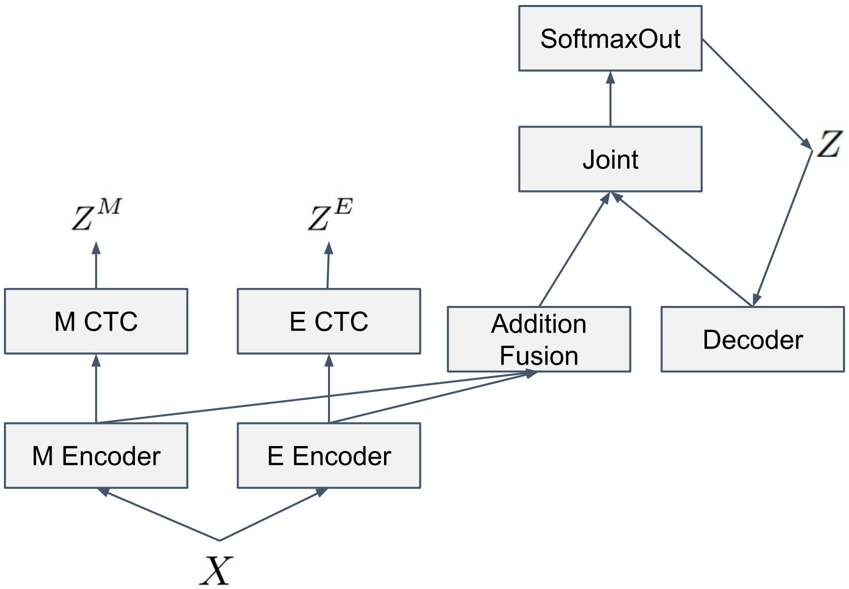
$$p(Z^E|X) \approx \prod_{t=1}^{T} p(z_t^E|X, \cancel{z_{1:t-1}^E})$$

Bilingual RNN-T Module:

$$p(Z|Z^M, Z^E) = \prod_{i=1}^{T+L} p(z_i|Z^M, Z^E, z_{1:i-1})$$

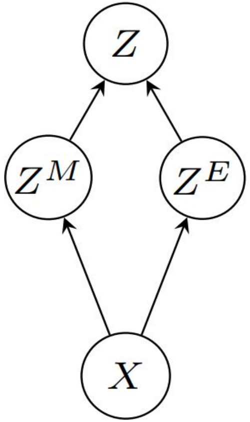
$$\begin{aligned}
\mathbf{h}_t^{\text{Enc}} &= \mathbf{h}_t^M + \mathbf{h}_t^E \\
\mathbf{h}_l^{\text{Dec}} &= \text{Decoder}(z_{1:l-1}) \\
\mathbf{h}_{t,l}^{\text{Jnt}} &= \text{Joint}(\mathbf{h}_t^{\text{Enc}}, \mathbf{h}_l^{\text{Dec}}) \\
p(z_i|\mathbf{h}^M, \mathbf{h}^E, z_{1:i-1}) &= \text{SoftmaxOut}(\mathbf{h}_{t,l}^{\text{Jnt}})
\end{aligned}$$

Conditional RNN Transducer

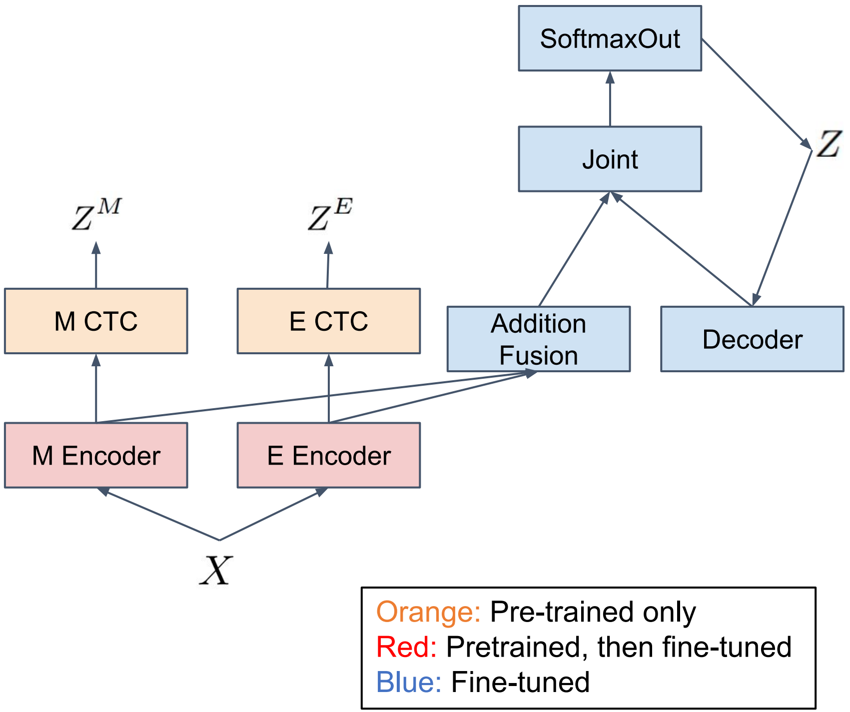


Full Network:

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



Conditional RNN Transducer



Implicit conditioning via pre-training:

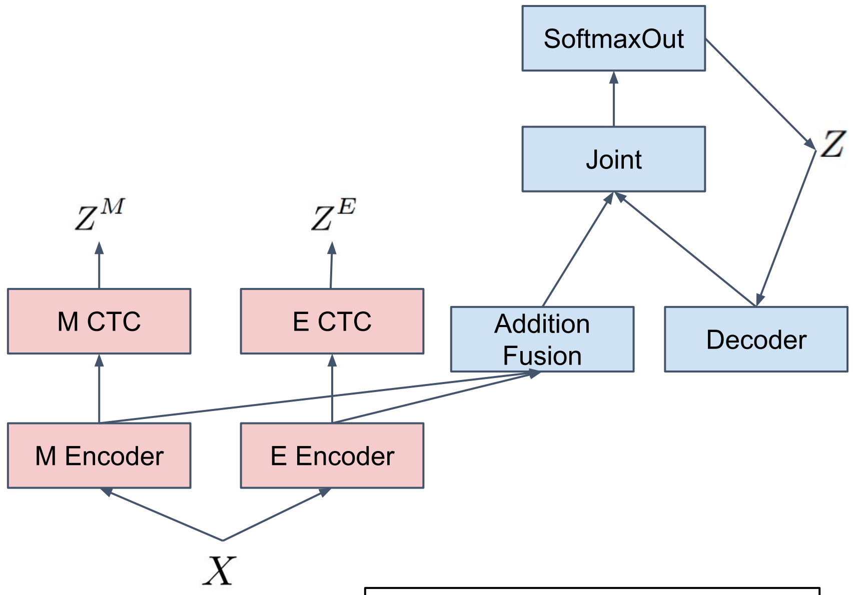
- Pre-train:

  $\mathcal{L}_{\text{M\_CTC}}$, $\mathcal{L}_{\text{E\_CTC}}$

- Fine-tune:

  $\lambda\mathcal{L}_{\text{RNNT}}$

Conditional RNN Transducer



Orange: Pre-trained only
Red: Pretrained, then fine-tuned
Blue: Fine-tuned

Explicit conditioning via masking:

- Pre-train:

  $\mathcal{L}_{\text{M\_CTC}}$, $\mathcal{L}_{\text{E\_CTC}}$

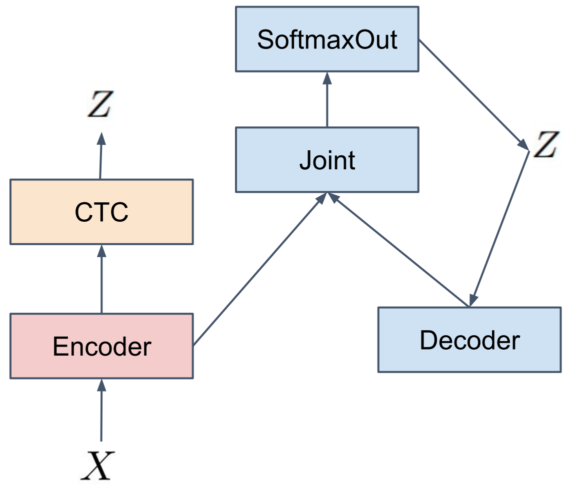
- Fine-tune:

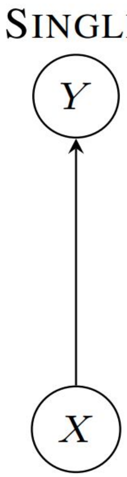
  $$\mathcal{L}_{\text{LS}} = \lambda\mathcal{L}_{\text{RNNT}} + (1-\lambda)(\mathcal{L}_{\text{M\_CTC}} + \mathcal{L}_{\text{E\_CTC}})$$

- Monolingual ground truths are obtained via language-specific masking of the bilingual ground truth → referred to as **Language-Separation**

Original Bilingual g.t.	什么是 Code-Switching
Masked Mandarin g.t.	什么是 <en>
Masked English g.t.	<zh> Code-Switching

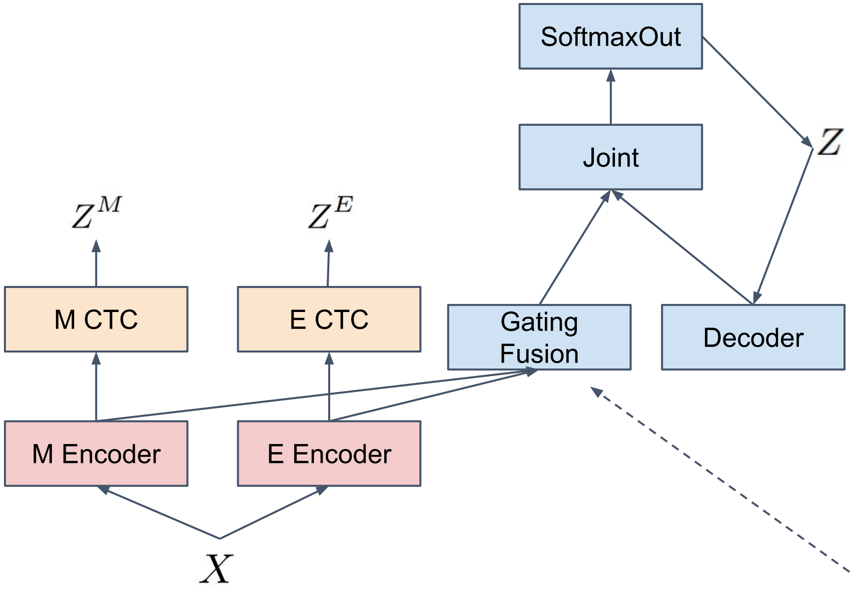
Single RNN-T Baseline

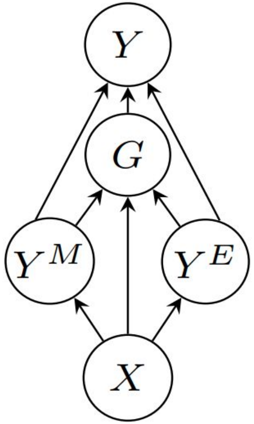




Orange: Pre-trained only
Red: Pretrained, then fine-tuned
Blue: Fine-tuned

Gating RNN-T Baseline





Orange: Pre-trained only
Red: Pretrained, then fine-tuned
Blue: Fine-tuned

$$\alpha_{\text{enc}} = \text{sigmoid}(w_\alpha(\tanh(w_{\text{enc}}^{\text{man}}(h^{\text{man}}) + w_{\text{enc}}^{\text{en}}(h^{\text{en}}))))$$

$$h = \alpha_{\text{enc}} * h^{\text{man}} + (1 - \alpha_{\text{enc}}) * h^{\text{en}}$$

Experimental Setup

Data:

- **200h Mandarin-English CS data** from ASRU’19 challenge
- **500h monolingual Mandarin data** from ASRU’19 challenge
  - 200h subset used for fine-tuning
- **700h monolingual accented English data** from King-ASR-190
  - 200h subset used for fine-tuning

Evaluation:

- **CS set** as measured by Mixed Error Rate (MER)
- **Monolingual Mandarin set** as measured by Character Error Rate (CER)
- **Monolingual English set** as measured by Word Error Rate (WER)

Main Results

Model Type	Model Name	Pre-trained Encoder(s)	Fine-tuning Data	CODE-SWITCHED MER	CODE-SWITCHED CER	CODE-SWITCHED WER	MONO-MAN CER	MONO-ENG WER
Direct	Vanilla RNN-T [21, 24]	✓	CS	12.3	9.9	34.3	17.9	81.4
Mixture	Gating RNN-T [22, 24]	✓	CS	11.5	9.1	33.0	17.7	**78.3**
Conditional	Our Proposed Model	✓	CS	11.5	9.1	33.2	15.5	82.9
Conditional	+ Language-Separation (LS)	✓	CS	**11.1**	**8.7**	**32.7**	**15.3**	82.7
Direct	Single RNN-T [21, 24]	✓	CS + M	11.3	9.3	30.8	6.5	17.8
Mixture	Gating RNN-T [22, 24]	✓	CS + M	11.2	8.8	34.7	5.7	34.6
Conditional	Our Proposed Model	✓	CS + M	10.3	8.2	29.5	5.4	16.5
Conditional	+ Language-Separation (LS)	✓	CS + M	**10.2**	**8.1**	**29.2**	**5.3**	**16.3**

- All models perform significantly better on monolingual sets when using monolingual fine-tuning

Main Results

Model Type	Model Name	Pre-trained Encoder(s)	Fine-tuning Data	CODE-SWITCHED MER	CODE-SWITCHED CER	CODE-SWITCHED WER	MONO-MAN CER	MONO-ENG WER
Direct	Vanilla RNN-T [21, 24]	✓	CS	12.3	9.9	34.3	17.9	81.4
Mixture	Gating RNN-T [22, 24]	✓	CS	11.5	9.1	33.0	17.7	**78.3**
Conditional	Our Proposed Model	✓	CS	11.5	9.1	33.2	15.5	82.9
Conditional	+ Language-Separation (LS)	✓	CS	**11.1**	**8.7**	**32.7**	**15.3**	82.7
Direct	Single RNN-T [21, 24]	✓	CS + M	11.3	9.3	30.8	6.5	17.8
Mixture	Gating RNN-T [22, 24]	✓	CS + M	11.2	8.8	34.7	5.7	34.6
Conditional	Our Proposed Model	✓	CS + M	10.3	8.2	29.5	5.4	16.5
Conditional	+ Language-Separation (LS)	✓	CS + M	**10.2**	**8.1**	**29.2**	**5.3**	**16.3**

- All models perform significantly better on monolingual sets when using monolingual fine-tuning
- Gating RNN-T outperforms Single RNN-T on CS set, but not on monolingual English

Main Results

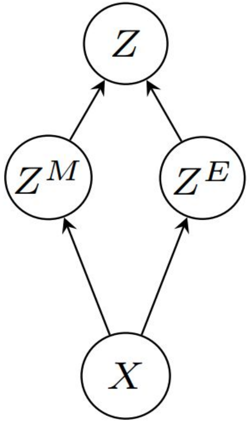
Model Type	Model Name	Pre-trained Encoder(s)	Fine-tuning Data	CODE-SWITCHED MER	CODE-SWITCHED CER	CODE-SWITCHED WER	MONO-MAN CER	MONO-ENG WER
Direct	Vanilla RNN-T [21, 24]	✓	CS	12.3	9.9	34.3	17.9	81.4
Mixture	Gating RNN-T [22, 24]	✓	CS	11.5	9.1	33.0	17.7	**78.3**
Conditional	Our Proposed Model	✓	CS	11.5	9.1	33.2	15.5	82.9
Conditional	+ Language-Separation (LS)	✓	CS	**11.1**	**8.7**	**32.7**	**15.3**	82.7
Direct	Single RNN-T [21, 24]	✓	CS + M	11.3	9.3	30.8	6.5	17.8
Mixture	Gating RNN-T [22, 24]	✓	CS + M	11.2	8.8	34.7	5.7	34.6
Conditional	Our Proposed Model	✓	CS + M	10.3	8.2	29.5	5.4	16.5
Conditional	+ Language-Separation (LS)	✓	CS + M	**10.2**	**8.1**	**29.2**	**5.3**	**16.3**

- All models perform significantly better on monolingual sets when using monolingual fine-tuning
- Gating RNN-T outperforms Single RNN-T on CS set, but not on monolingual English
- Conditional RNN-T outperforms both baselines across CS and monolingual sets

Analysis of Language-Separation (LS) Ability

Recall:

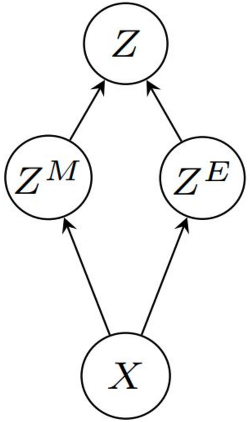
$$z_t = \begin{cases} z_t^M, & \text{if } z_t^M \in \mathcal{V}^M \text{ and } z_t^E = \varnothing \\ z_t^E, & \text{if } z_t^E \in \mathcal{V}^E \text{ and } z_t^M = \varnothing \\ \varnothing, & \text{if } z_t^M = \varnothing \text{ and } z_t^E = \varnothing \end{cases}$$



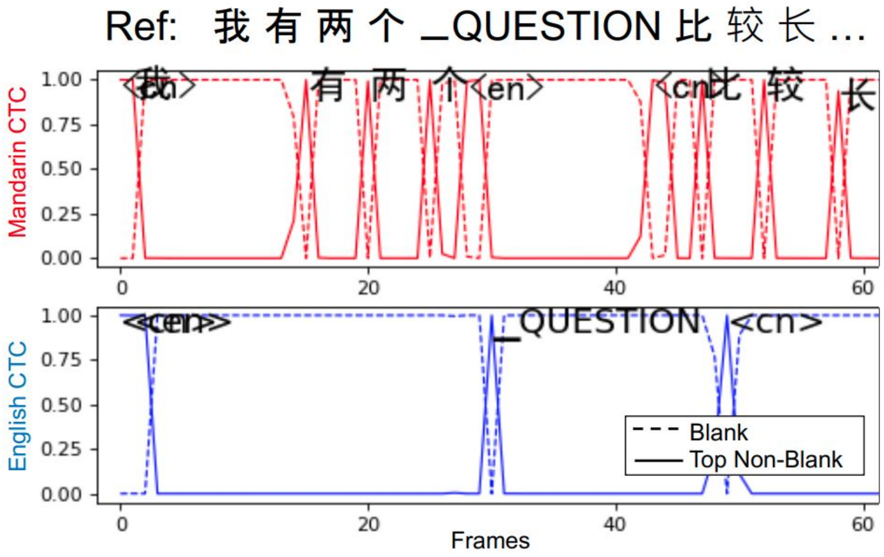
Analysis of Language-Separation (LS) Ability

Recall:

$$z_t = \begin{cases} z_t^M, & \text{if } z_t^M \in \mathcal{V}^M \text{ and } z_t^E = \varnothing \\ z_t^E, & \text{if } z_t^E \in \mathcal{V}^E \text{ and } z_t^M = \varnothing \\ \varnothing, & \text{if } z_t^M = \varnothing \text{ and } z_t^E = \varnothing \end{cases}$$



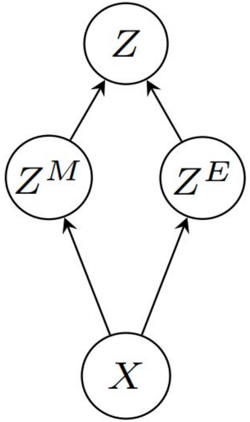
Visualizing Language Separation:



Analysis of Language-Separation (LS) Ability

Recall:

$$z_t = \begin{cases} z_t^M, & \text{if } z_t^M \in \mathcal{V}^M \text{ and } z_t^E = \varnothing \\ z_t^E, & \text{if } z_t^E \in \mathcal{V}^E \text{ and } z_t^M = \varnothing \\ \varnothing, & \text{if } z_t^M = \varnothing \text{ and } z_t^E = \varnothing \end{cases}$$



Evaluating the monolingual CTC sub-nets:

	Man Portion of CS			Eng Portion of CS		
Model	Sub-Net	CER	INS	Sub-Net	WER	INS
Cond. RNN-T	$p(Z^M\|X)$	11.8	3.7	$p(Z^E\|X)$	42.7	7.9
Cond. RNN-T + LS	$p(Z^M\|X)$	**8.6**	**0.7**	$p(Z^E\|X)$	**37.1**	**4.6**

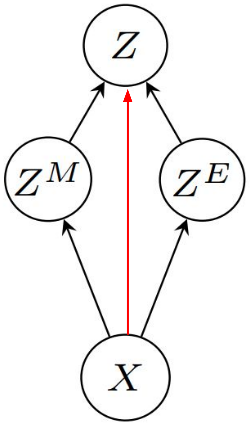
- We evaluate the monolingual CTC outputs against the language-specific portions of the CS reference
- Both models can perform reasonable language diarization
- Conditional RNN-T + LS has reduced insertion errors

Conditional Independence of Bilingual Module

Recall:

$$
\begin{aligned}
p(Z|X) &= p(Z, Z^M, Z^E|X) \\
&= p(Z|Z^M, Z^E, X)p(Z^M, Z^E|X) \\
&\approx p(Z|Z^M, Z^E, \cancel{X})p(Z^M|X)p(Z^E|X)
\end{aligned}
$$

Conditional independence

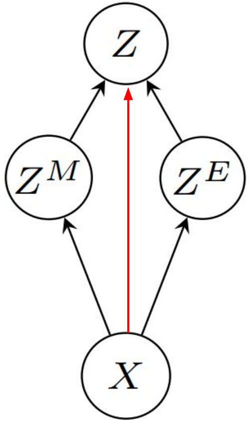


Conditional Independence of Bilingual Module

Recall:

$$
\begin{aligned}
p(Z|X) &= p(Z, Z^M, Z^E|X) \\
&= p(Z|Z^M, Z^E, X)p(Z^M, Z^E|X) \\
&\approx p(Z|Z^M, Z^E, \cancel{X})p(Z^M|X)p(Z^E|X)
\end{aligned}
$$

Conditional independence

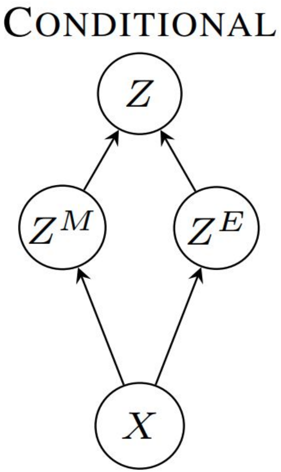


Experimental Validation:

Model	Bilingual Condition	CODE-SWITCHED		
		MER	CER	WER
Cond. RNN-T + LS	$p(Z\|Z^M, Z^E)$	**11.1**	**8.9**	**31.1**
3-Enc. RNN-T + LS	$p(Z\|Z^M, Z^E, X)$	11.2	9.0	**31.1**

- The 3-encoder variant removes the conditional independence assumption by directly encoding speech features, *X*, to the bilingual module
- This dependency adds no additional information as the monolingual alignments are enough to determine the bilingual alignment

Recap: Did we satisfy our desiderata?



- ☑ Can we build CS + bilingual ASR with **monolingual sub-components**…
- ☑ …where the final output is **conditioned only on those 2 sub-components** and nothing else?
- ☑ And does such a conditional approach more **efficiently leverage monolingual and CS training data**?

Thank You!

Session: SPE-14: Multi-lingual ASR
Session Time: Sunday, 8 May, 23:00 - 23:45 (Singapore Time, UTC +8)

Brian Yan[1], *Chunlei Zhang*[2], *Meng Yu*[2], *Shi-Xiong Zhang*[2], *Siddharth Dalmia*[1], *Dan Berrebbi*[1], *Chao Weng*[3], *Shinji Watanabe*[1], *Dong Yu*[2]

[1]Carnegie Mellon University, USA, [2]Tencent AI Lab, USA, [3]Tencent AI Lab, China

byan@cs.cmu.edu