CMU's IWSLT 2022 Dialect Speech Translation System

Brian Yan[1] **Patrick Fernandes**[1,2] **Siddharth Dalmia**[1] **Jiatong Shi**[1]
Yifan Peng[3] **Dan Berrebbi**[1] **Xinyi Wang**[1] **Graham Neubig**[1] **Shinji Watanabe**[1,4]
[1]Language Technologies Institute, Carnegie Mellon University, USA
[2]Instituto Superior Técnico & LUMLIS (Lisbon ELLIS Unit), Portugal
[3]Electrical and Computer Engineering, Carnegie Mellon University, USA
[4]Human Language Technology Center of Excellence, Johns Hopkins University, USA
{byan, pfernand, sdalmia, jiatongs}@cs.cmu.edu

Day1	May 26, 2022
Time (Dublin)	**Session**
15:30-17:30	Poster Session: System Papers

Carnegie Mellon University
Language Technologies Institute

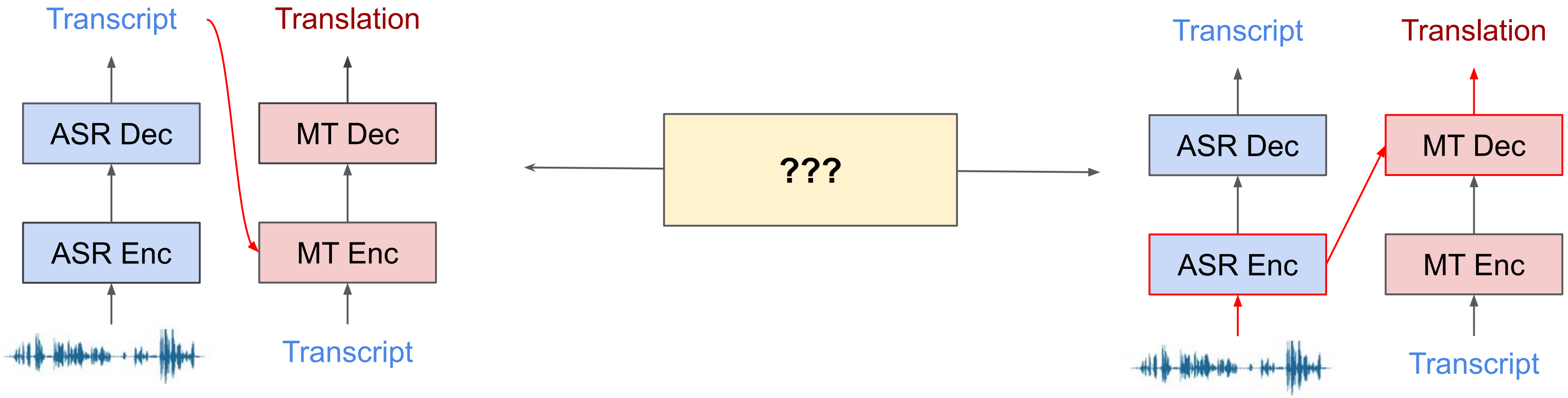
ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

Image Source: Salesky & Black https://www.aclweb.org/anthology/2020.acl-main.217.pdf

The End-to-End Fallacy

Fully Cascaded

Fully Direct
(w/ multi-tasking)



Con: early error propagation

Con: sensitive to noisy ASR transcription

Pro: relatively tons of ASR and MT data

Pro: post-processing (e.g. ROVER) / external models for ASR

Pro: no intermediate representation

Pro: no intermediate representation

Con: less ST data + less data efficient with ASR & MT pre-training multi-task (wasted subnets)

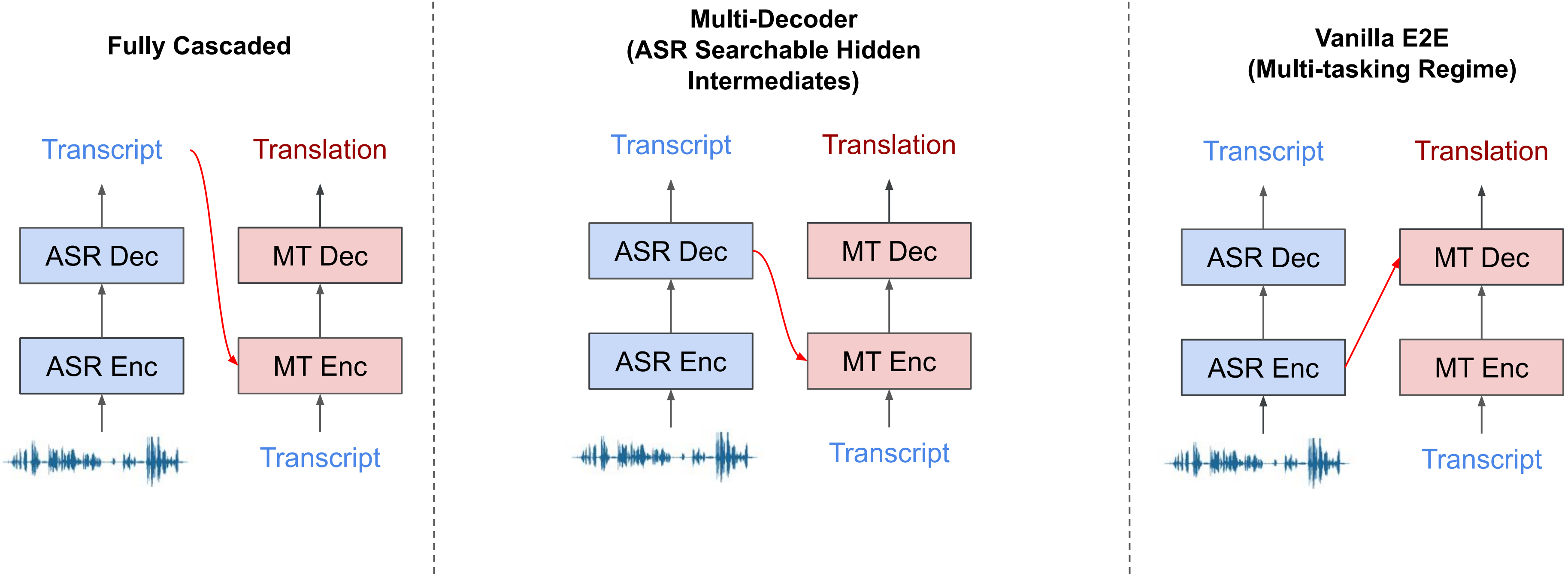
Con: single retrieval stage

Hybrid Approaches to Speech Translation

1. Multi-Decoder with Searchable Hidden Intermediates

2. Minimum Bayes-Risk Decoding

Multi-Decoder vs. Cascade vs. Vanilla E2E

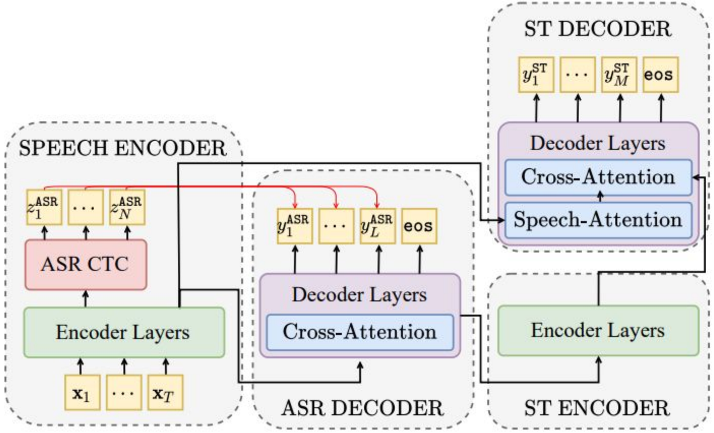


- ❏ Beam search over ASR output w/ use of external models (e.g. LM, CTC)
- ❏ Use of post-processing to improve ASR output (e.g. ROVER)

Siddharth Dalmia, Brian Yan, Vikas Raunak, Florian Metze, and Shinji Watanabe, “Searchable Hidden Intermediates for End-to-End Models of Decomposable Sequence Tasks,” Proc. NAACL’21 (link)

Multi-Decoder with Searchable Hidden Intermediates

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{CE}}^{\text{ASR}} + \lambda_2 \mathcal{L}_{\text{CTC}}^{\text{ASR}} + \lambda_3 \mathcal{L}_{\text{CE}}^{\text{ST}}$$



(a) Multi-Decoder

+Searchable Hidden Intermediates: ASR decoder representations are retrieved (e.g. via beam search) and passed to the ST Encoder

---

Algorithm 1 Beam Search for Hidden Intermediates: We perform beam search to approximate the most likely sequence for the sub-task $\mathcal{A} \rightarrow \mathcal{B}$, $\mathbf{y}_{\text{BEAM}}^{\mathcal{B}}$, while collecting the corresponding $\text{DECODER}_{\mathcal{B}}$ hidden representations, $\mathbf{h}_{\text{BEAM}}^{D_{\mathcal{B}}}$. The output $\mathbf{h}_{\text{BEAM}}^{D_{\mathcal{B}}}$, is passed to the final sub-network to predict final output $\mathcal{C}$ and $\mathbf{y}_{\text{BEAM}}^{\mathcal{B}}$ is used for monitoring performance on predicting $\mathcal{B}$.

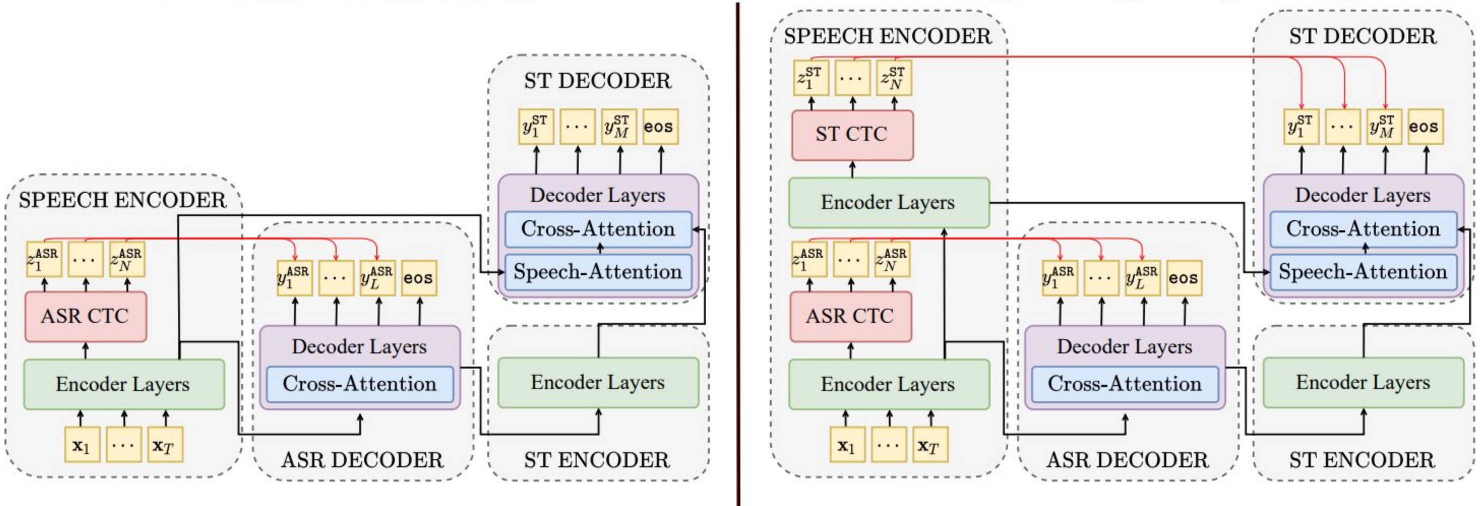
Initialize: $\text{BEAM} \leftarrow \{\text{sos}\}$; $\text{k} \leftarrow$ beam size;
$\mathbf{h}^{E_{\mathcal{A}}} \leftarrow \text{ENCODER}_{\mathcal{A}}(\mathbf{x})$
for l=1 **to** $\text{max}_{\text{STEPS}}$ **do**
  for $\mathbf{y}_{l-1}^{\mathcal{B}} \in \text{BEAM}$ **do**
    $\mathbf{h}_l^{D_{\mathcal{B}}} \leftarrow \text{DECODER}_{\mathcal{B}}(\mathbf{h}^{E_{\mathcal{A}}}, \mathbf{y}_{l-1}^{\mathcal{B}})$
    for $\mathbf{y}_l^{\mathcal{B}} \in \mathbf{y}_{l-1}^{\mathcal{B}} + \{\mathcal{V}\}$ **do**
      $s_l \leftarrow P_{\mathcal{A}\rightarrow\mathcal{B}}(\mathbf{y}_l^{\mathcal{B}} \mid \mathbf{x})^{1\text{-}\lambda} P_{\text{EXT}}(\mathbf{y}_l^{\mathcal{B}})^{\lambda}$
      $\mathcal{H} \leftarrow (s_l, \mathbf{y}_l^{\mathcal{B}}, \mathbf{h}_l^{D_{\mathcal{B}}})$
    end for
  end for
  $\text{BEAM} \leftarrow \text{arg}^{\text{k}}\text{max}(\mathcal{H})$
end for
$(s^{\mathcal{B}}, \mathbf{y}_{\text{BEAM}}^{\mathcal{B}}, \mathbf{h}_{\text{BEAM}}^{D_{\mathcal{B}}}) \leftarrow \text{argmax}(\text{BEAM})$
Return $\mathbf{y}_{\text{BEAM}}^{\mathcal{B}} \rightarrow \text{SUB}_{\mathcal{A}\rightarrow\mathcal{B}}\text{NET}$ Monitoring
Return $\mathbf{h}_{\text{BEAM}}^{D_{\mathcal{B}}} \rightarrow$ Final $\text{SUB}_{\mathcal{B}\rightarrow\mathcal{C}}\text{NET}$

Siddharth Dalmia, Brian Yan, Vikas Raunak, Florian Metze, and Shinji Watanabe, "Searchable Hidden Intermediates for End-to-End Models of Decomposable Sequence Tasks," Proc. NAACL'21 (link)

Multi-Decoder with Searchable Hidden Intermediates

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{CE}}^{\text{ASR}} + \lambda_2 \mathcal{L}_{\text{CTC}}^{\text{ASR}} + \lambda_3 \mathcal{L}_{\text{CE}}^{\text{ST}}$$

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{CE}}^{\text{ASR}} + \lambda_2 \mathcal{L}_{\text{CTC}}^{\text{ASR}} + \lambda_3 \mathcal{L}_{\text{CE}}^{\text{ST}} + \lambda_4 \mathcal{L}_{\text{CTC}}^{\text{ST}}$$



(a) Multi-Decoder

(b) Multi-Decoder w/ Hierarchical Encoder + CTC/Attn ST Decoding

+Searchable Hidden Intermediates: ASR decoder representations are retrieved (e.g. via beam search) and passed to the ST Encoder

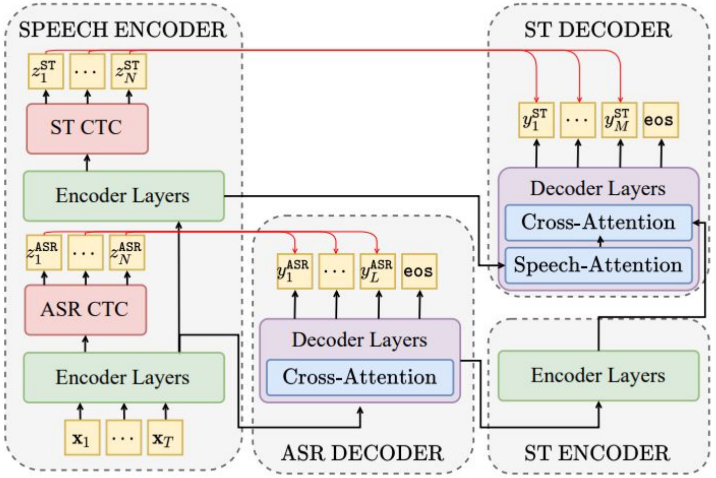
+Hierarchical Encoder: re-orders speech encoder
+Joint CTC/Attn ST Decoding: length normalization
+ASR CTC Sampling: simulates ASR errors in training

Multi-Decoder with Searchable Hidden Intermediates

	test1
Model Name	BLEU(↑)
Encoder-Decoder	16.0
Multi-Decoder	17.1
+ ASR CTC Sampling	17.6
+ Hierarchical Encoder	17.9
+ Joint CTC/Attn ST Decoding (D4)	18.2
+ ASR CTC Sampling	**18.4**

+2.4 BLEU

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{CE}}^{\text{ASR}} + \lambda_2 \mathcal{L}_{\text{CTC}}^{\text{ASR}} + \lambda_3 \mathcal{L}_{\text{CE}}^{\text{ST}} + \lambda_4 \mathcal{L}_{\text{CTC}}^{\text{ST}}$$



(b) Multi-Decoder w/ Hierarchical Encoder + CTC/Attn ST Decoding

+Hierarchical Encoder: re-orders speech encoder
+Joint CTC/Attn ST Decoding: length normalization
+ASR CTC Sampling: simulates ASR errors in training

Siddharth Dalmia, Brian Yan, Vikas Raunak, Florian Metze, and Shinji Watanabe, “Searchable Hidden Intermediates for End-to-End Models of Decomposable Sequence Tasks,” Proc. NAACL’21 (link)

Guiding Multi-Decoder Representations

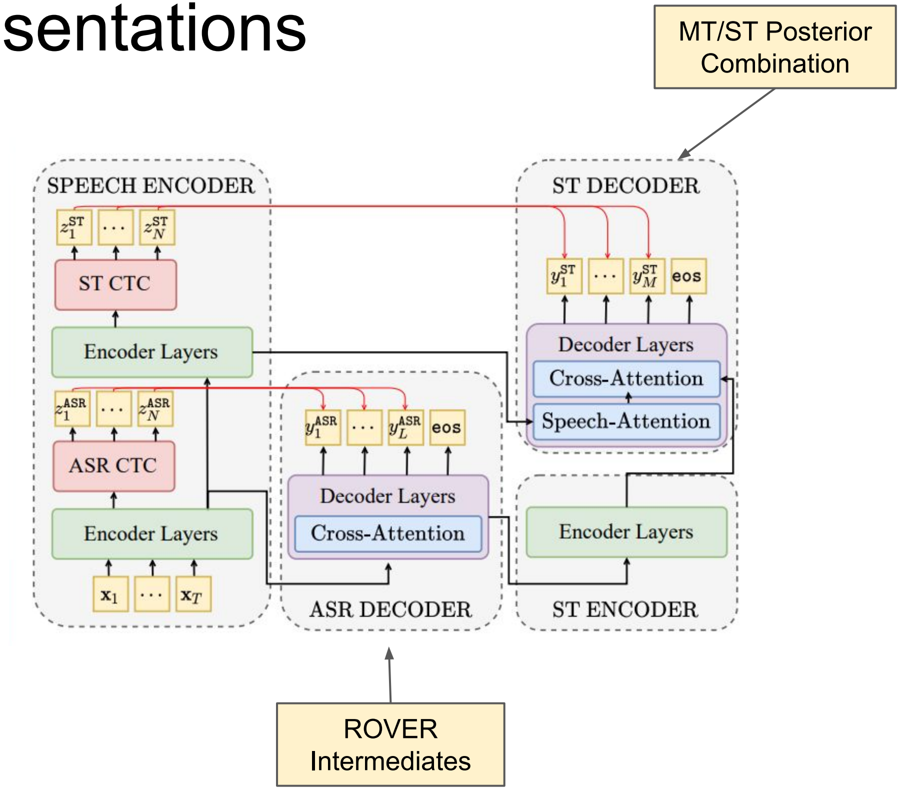
ASR Decoder:

- We retrieve the hidden representations for ASR outputs generated by ROVER combination
- No intermediate beam search is required

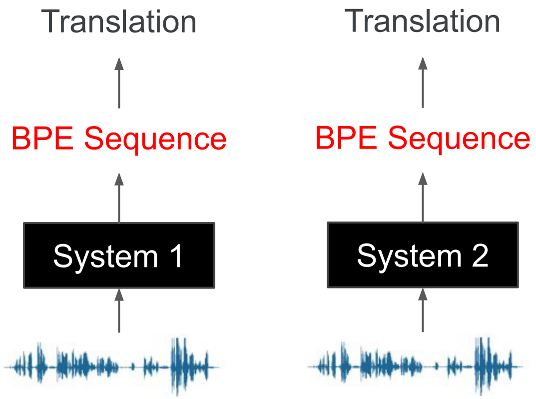
ST Decoder:

- External MT/ST models are used for posterior combination, along with joint CTC/Attn. ST decoding

*we do not consider these Multi-Decoder variants to be purely E2E

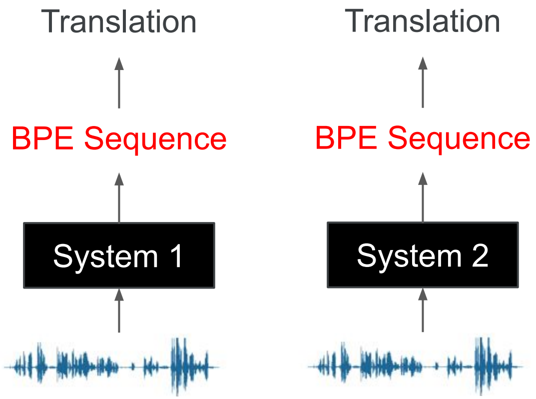


Minimum Bayes-Risk Decoding



- Posterior comb. distinct BPE vocabularies?
- ROVER-like align-then-vote - fit for translation?

Minimum Bayes-Risk Decoding



$$\hat{y}_{\text{MBR}} = \underset{y\in\bar{\mathcal{Y}}_{\text{cands}}}{\arg\max}\ \underbrace{\mathbb{E}_{Y\sim p_\theta(y|x)}[u(Y,y)]}_{\approx\ \frac{1}{M}\sum_{j=1}^{M} u(y^{(j)},y)}\ ,$$

Candidate set

BLEU metric

Sample set

- Posterior comb. distinct BPE vocabularies?
- ROVER-like align-then-vote - fit for translation?

- Both systems generate candidates and samples
- Risk of each candidate is measured as the avg. BLEU against all of the samples as references
- Systems can be black boxes

Minimum Bayes-Risk Decoding

ID	Type	Model Name	Child System(s)	Dialect Transfer	test1 BLEU(↑)	test2 BLEU(↑)
C1	Cascade	ASR Mixing Cascade	A1,B1	✗	16.4	-
C2	Cascade	+ ASR Rover Comb.	A2,B1	✗	16.7	-
C3	Cascade	+ MT Posterior Comb.	A2,B2	✗	17.5	18.6
C4	Cascade	ASR Mixing Cascade	A3,B3	✓	17.3	-
C5	Cascade	+ ASR Rover Comb.	A4,B3	✓	17.4	-
C6	Cascade	+ MT Posterior Comb.	A4,B4	✓	**17.9**	**19.4**
D1	E2E ST	Hybrid Multi-Decoder	-	✗	17.7	-
D2	Mix	+ ROVER Intermediates	A2	✗	18.1	19.1
D3	Mix	+ ST/MT Posterior Comb.	A2,B5	✗	18.7	19.7
D4	E2E ST	Hybrid Multi-Decoder	-	✓	18.2	-
D5	Mix	+ ROVER Intermediates	A4	✓	18.3	19.5
D6	Mix	+ ST/MT Posterior Comb.	A4,B5	✓	**18.9**	**19.8**
E1	Mix	Min. Bayes-Risk Ensemble	C3,D3	✗	19.2	20.4
E2	Mix	Min. Bayes-Risk Ensemble	C6,D6	✓	**19.5**	**20.8**

+1.3 BLEU

+0.6/+1.0 BLEU

Thanks!

CMU's IWSLT 2022 Dialect Speech Translation System

Brian Yan[1] **Patrick Fernandes**[1,2] **Siddharth Dalmia**[1] **Jiatong Shi**[1]
Yifan Peng[3] **Dan Berrebbi**[1] **Xinyi Wang**[1] **Graham Neubig**[1] **Shinji Watanabe**[1,4]
[1]Language Technologies Institute, Carnegie Mellon University, USA
[2]Instituto Superior Técnico & LUMLIS (Lisbon ELLIS Unit), Portugal
[3]Electrical and Computer Engineering, Carnegie Mellon University, USA
[4]Human Language Technology Center of Excellence, Johns Hopkins University, USA
`{byan, pfernand, sdalmia, jiatongs}@cs.cmu.edu`