Dual-decoder Transformer for Joint Automatic Speech Recognition and Multilingual Speech Translation

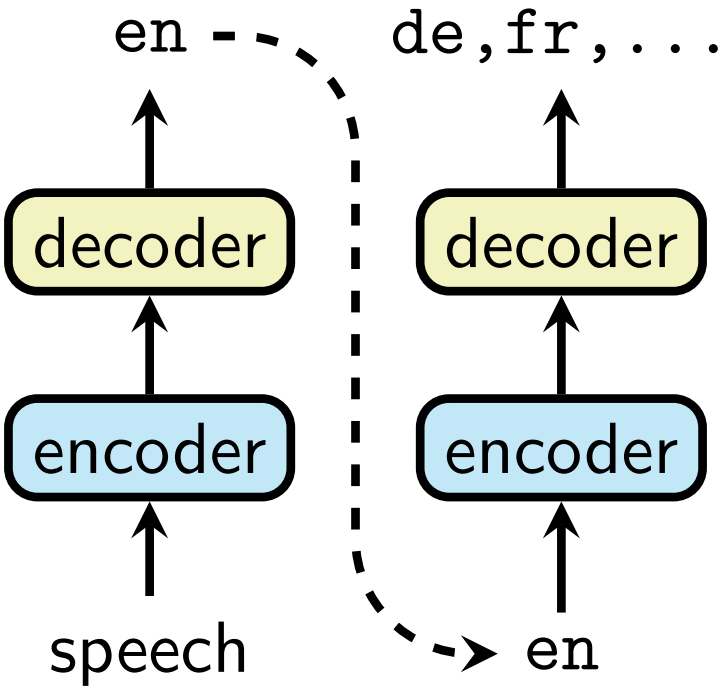
Hang Le[1] Juan Pino[2] Changhan Wang[2]
Jiatao Gu[2] Didier Schwab[1] Laurent Besacier[1]

[1]Univ. Grenoble Alpes, CNRS, LIG [2]Facebook AI

FACEBOOK AI

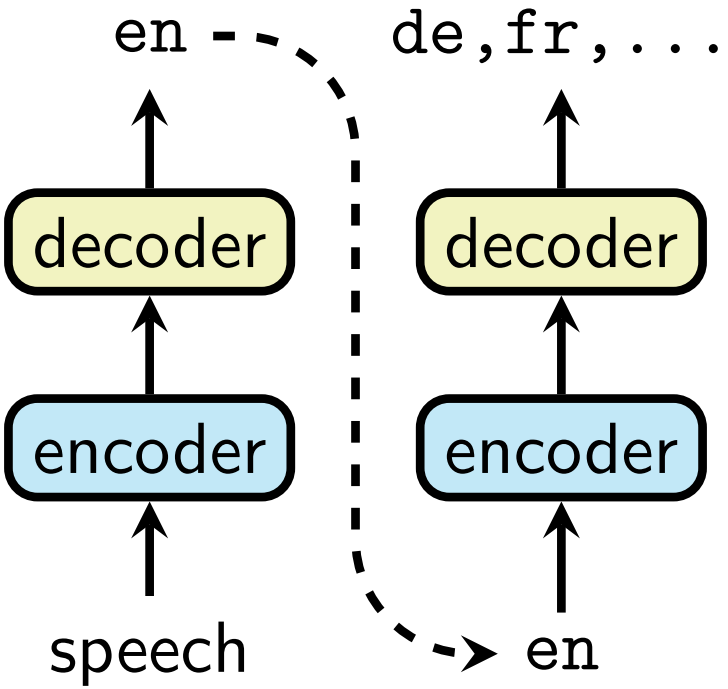
Outline

Cascade models



(Stentiford and Steer, 1988; Waibel et al., 1991)

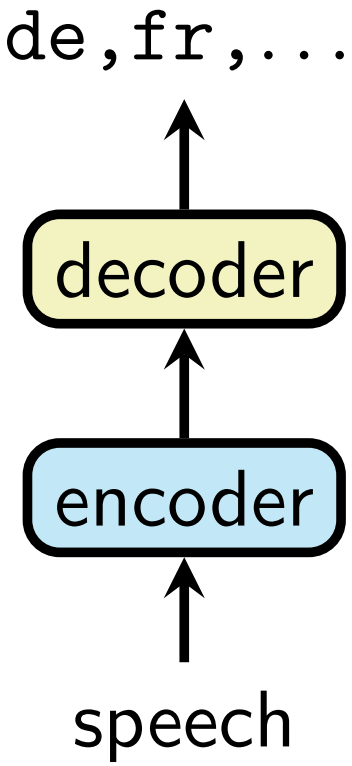
Cascade models



(Stentiford and Steer, 1988; Waibel et al., 1991)

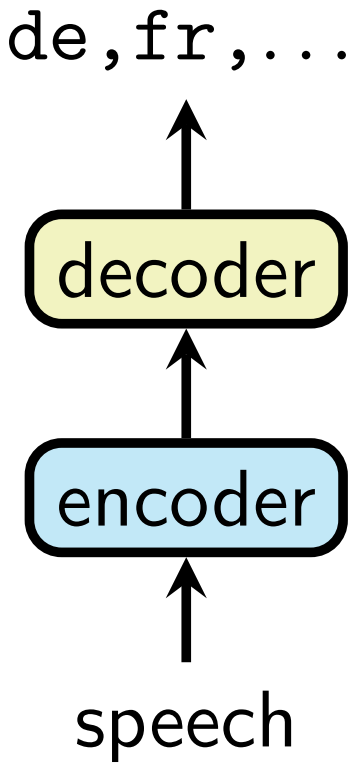
+ Strong performance.
+ Leverage ASR and MT datasets.
– Not end-to-end trainable $\rightarrow$ error propagation.
– High latency.

End-to-end models: bypassing ASR



(Bérard et al., 2016; Weiss et al., 2017; Bérard et al., 2018)

End-to-end models: bypassing ASR



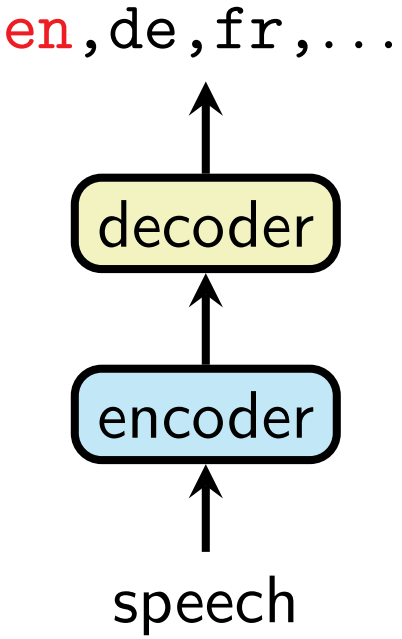
(Bérard et al., 2016; Weiss et al., 2017; Bérard et al., 2018)

- \+ End-to-end trainable $\rightarrow$ Reduce error propagation.
- \+ Low latency.
- – Transcripts not available (they may be beneficial to users).

End-to-end models: with ASR

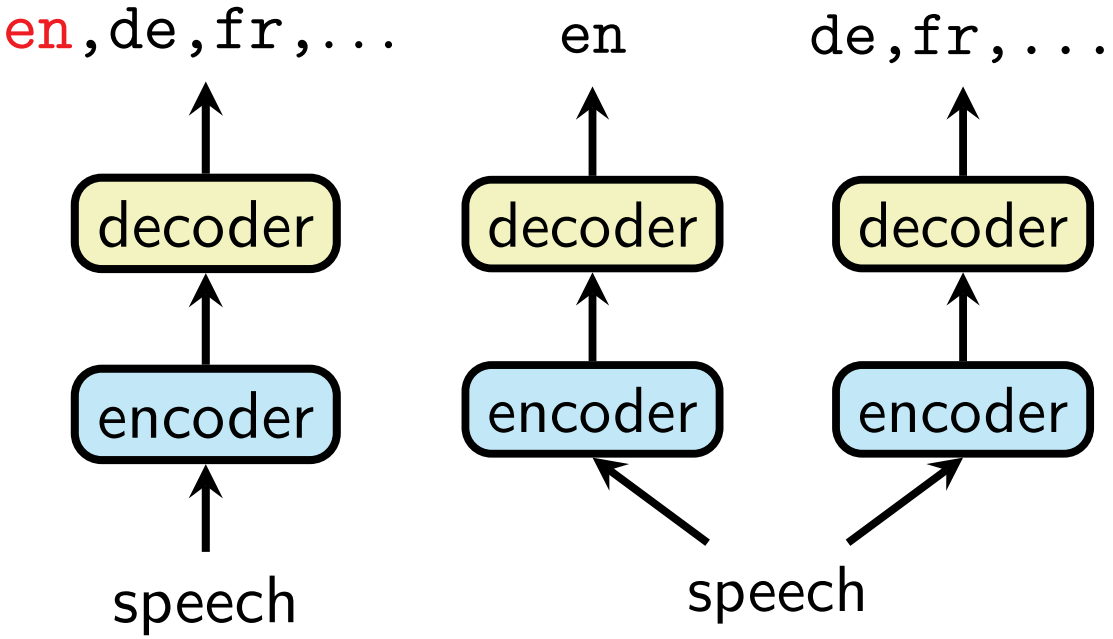
- Display of *transcripts alongside translations* can be *useful* in many applications (Sperber et al., 2020).
- *Transcripts* can *improve* translation *performance* (Gangi et al., 2019).
- *End-to-end models* featuring a *coupled inference procedure* are able to achieve *strong consistency* (Sperber et al., 2020).

End-to-end models: with ASR



en-as-additional-language
(Gangi et al., 2019.)

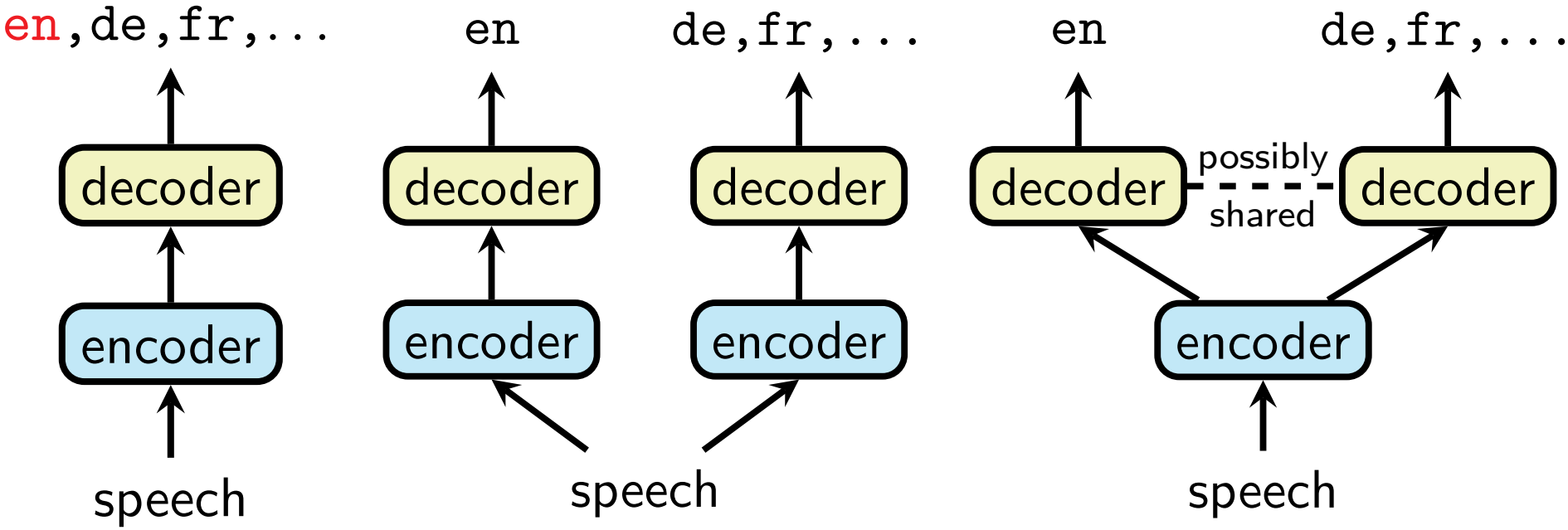
End-to-end models: with ASR



en-as-additional-language
(Gangi et al., 2019.)

Independent
(Sperber et al., 2020)

End-to-end models: with ASR

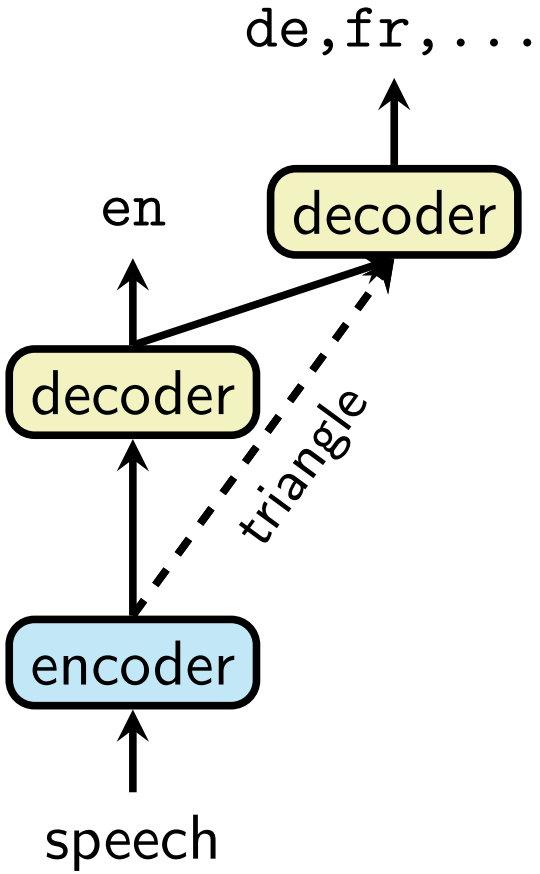


en-as-additional-language
(Gangi et al., 2019.)

Independent
(Sperber et al., 2020)

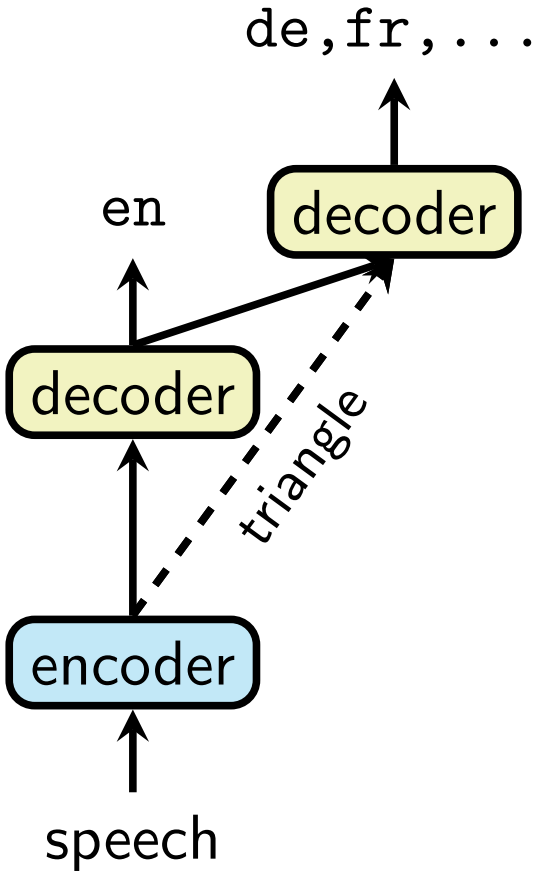
Multitask (Anastasopoulos and Chiang 2018, Sperber et al., 2020)
Shared (Sperber et al., 2020)

End-to-end models: with ASR

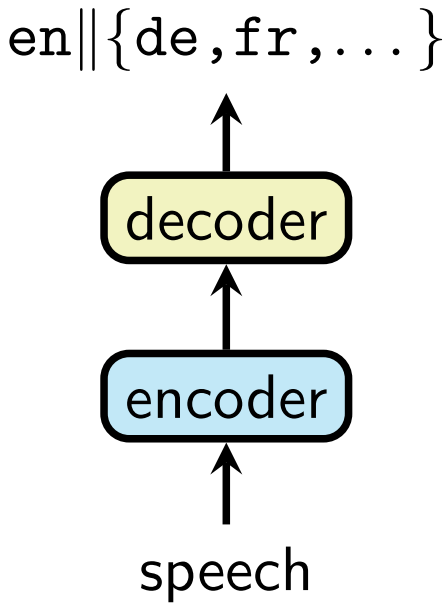


Two-stage, Triangle
(Anastasopoulos and Chiang, 2018; Sperber et al., 2019; Sperber et al., 2020)

End-to-end models: with ASR

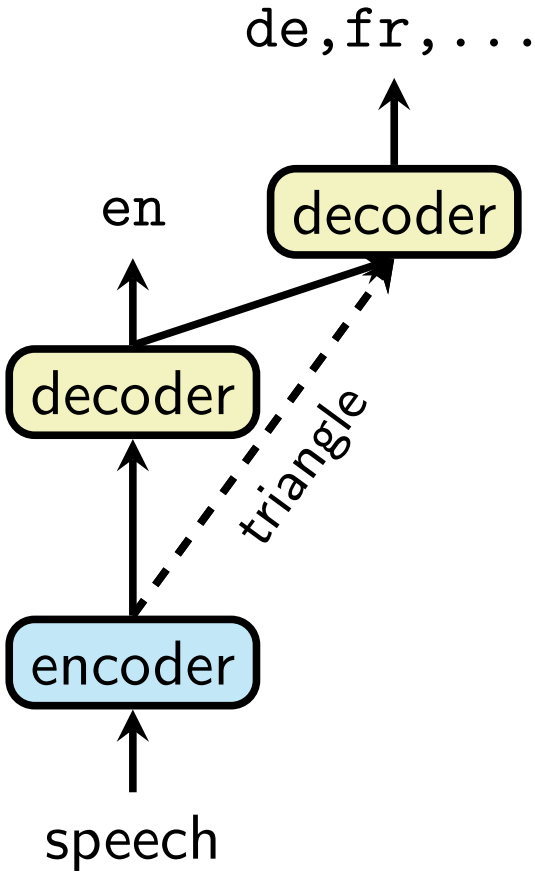


Two-stage, Triangle
(Anastasopoulos and Chiang, 2018; Sperber et al., 2019; Sperber et al., 2020)

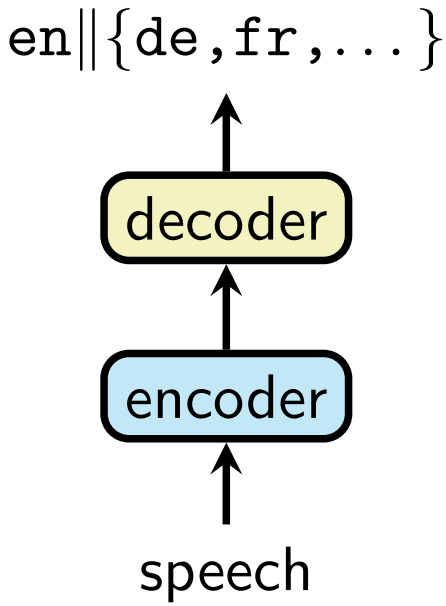


Concatenated
(Sperber et al., 2020)

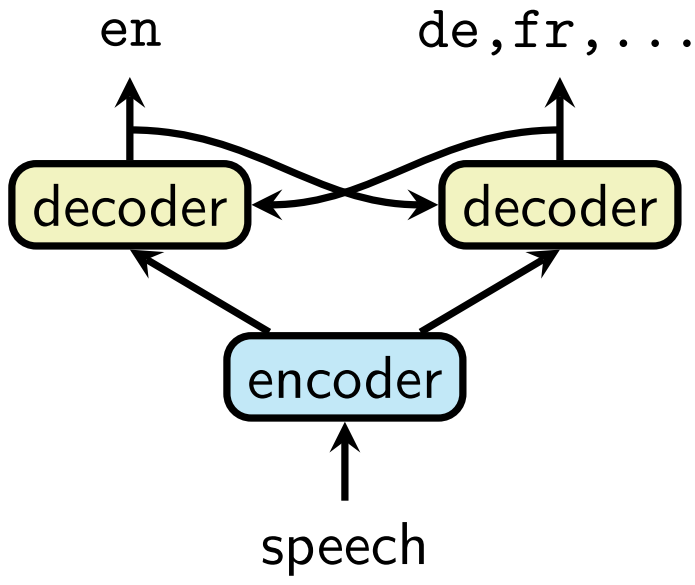
End-to-end models: with ASR



Two-stage, Triangle
(Anastasopoulos and Chiang, 2018; Sperber et al., 2019; Sperber et al., 2020)



Concatenated
(Sperber et al., 2020)



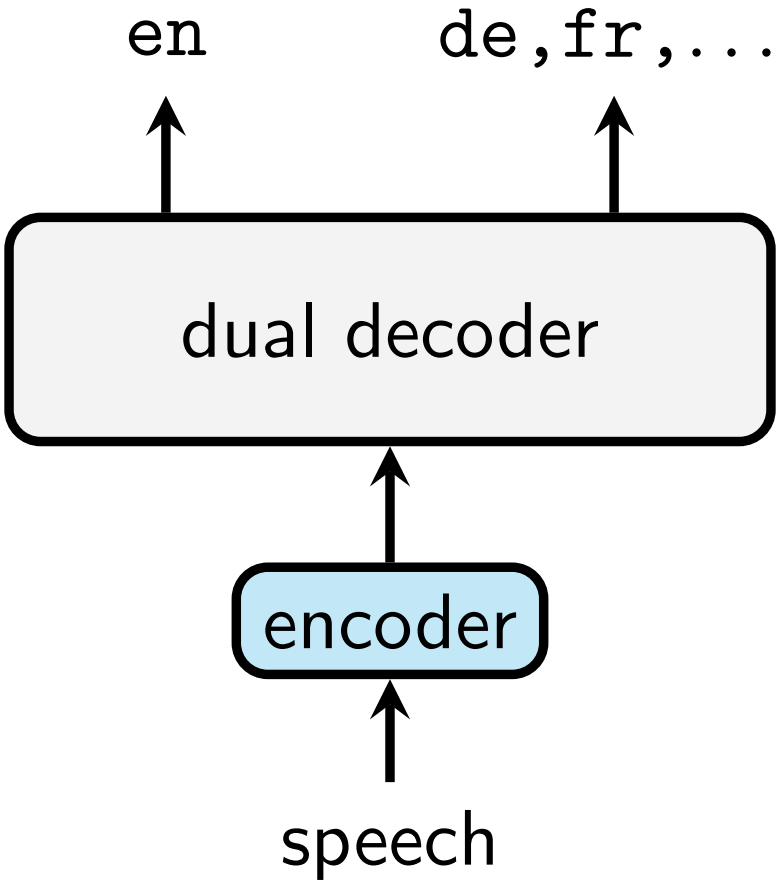
Interactive decoding
(Liu et al., 2020)

Dual-decoder Transformer

- Motivated by previous work, but *more general*.
- *Flexible*: level of *interaction between decoders* is a design choice.

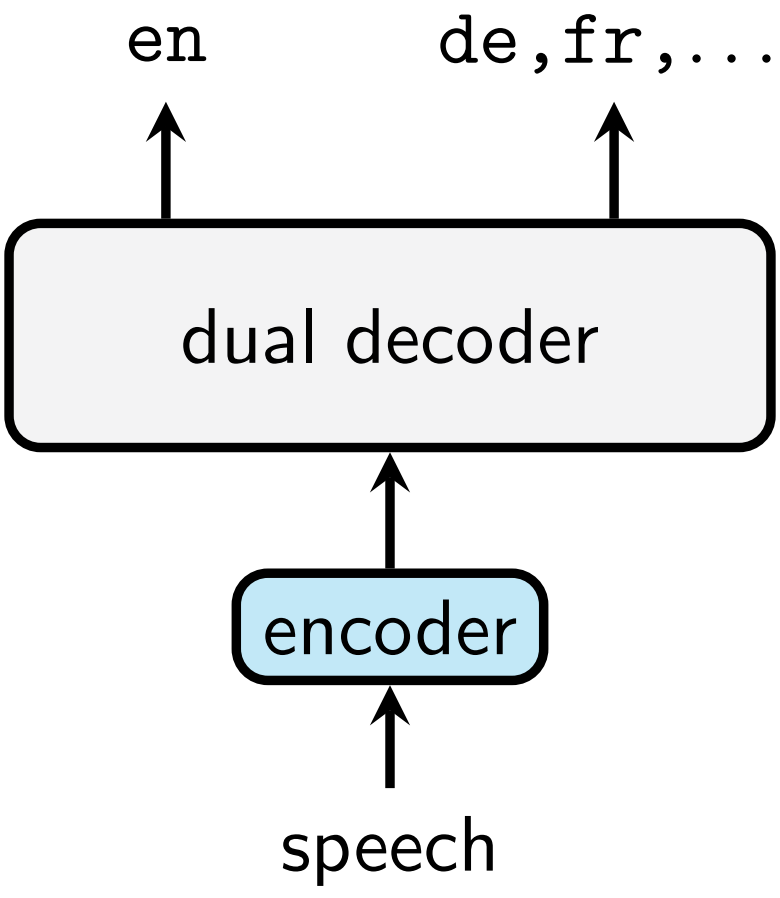
Dual-decoder Transformer

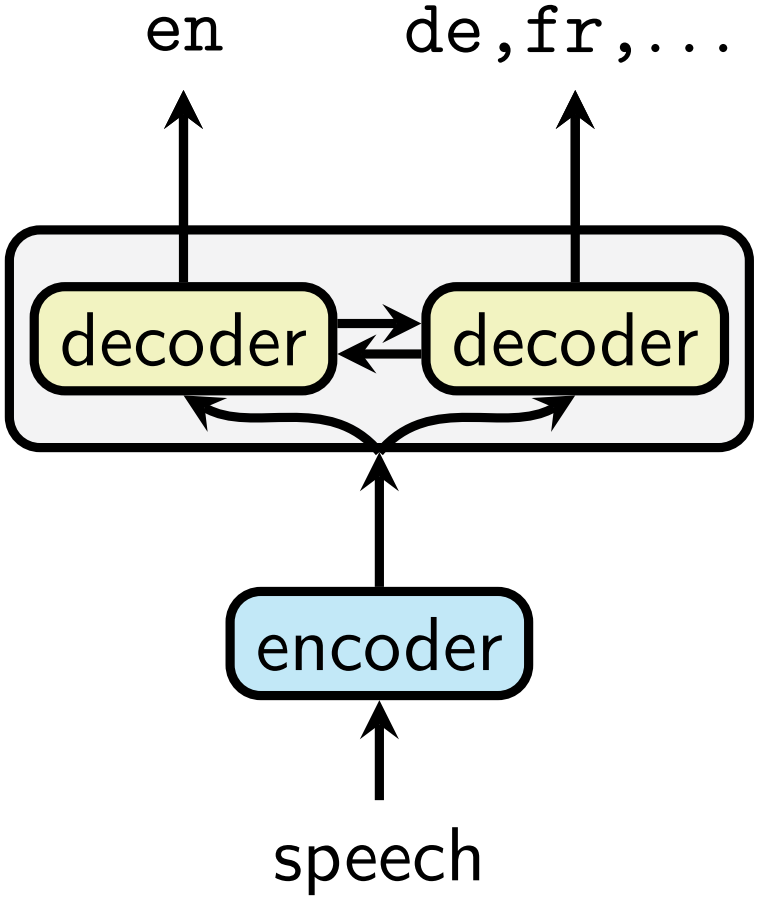
- Motivated by previous work, but *more general*.
- *Flexible*: level of *interaction between decoders* is a design choice.



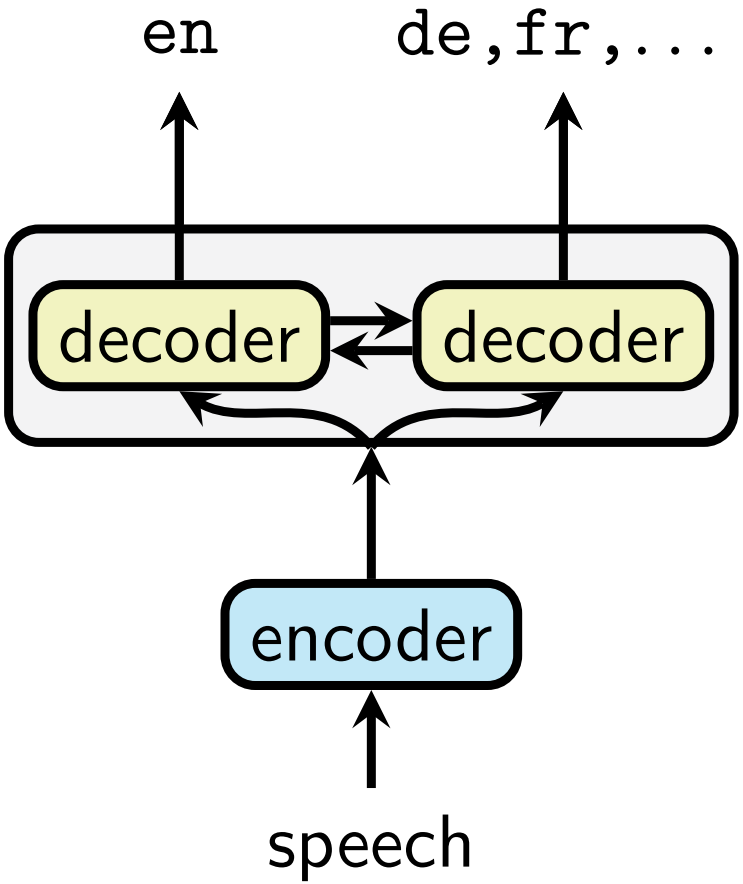
Dual-decoder Transformer

- Motivated by previous work, but *more general*.
- *Flexible*: level of *interaction between decoders* is a design choice.



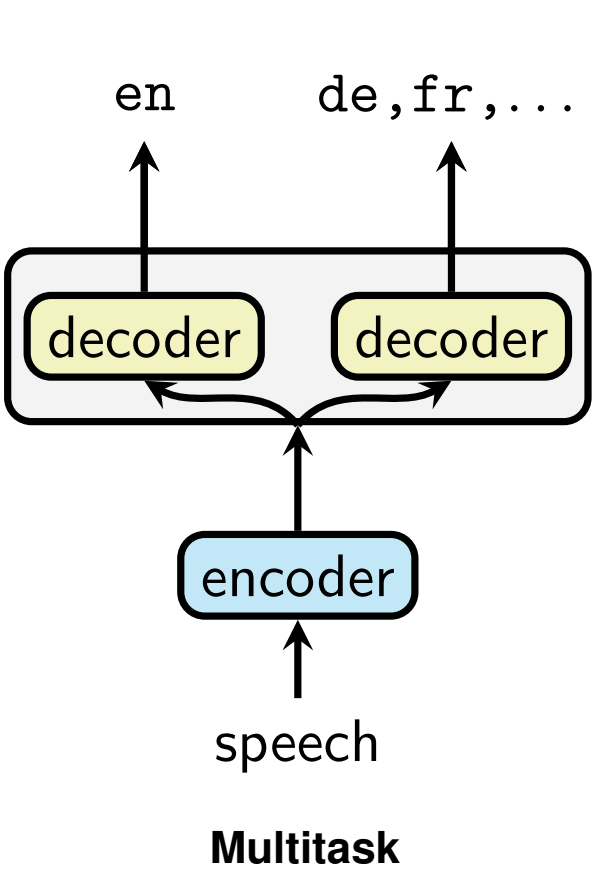


Dual-decoder Transformer

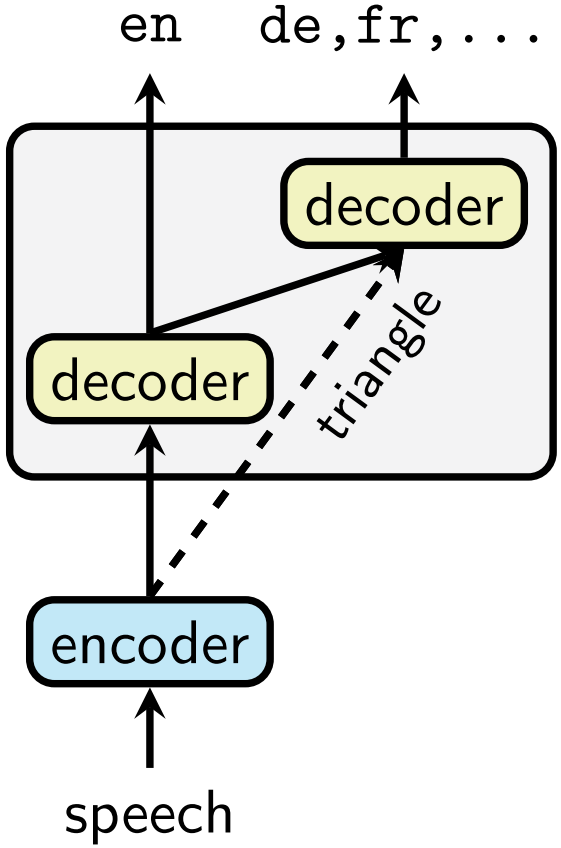


- + *Synchronous generation* of transcripts and translations.
- + *Generality*: several existing models are special cases (next slide).
- + ASR and ST tasks are *complementary* $\rightarrow$ *Strong performance*.

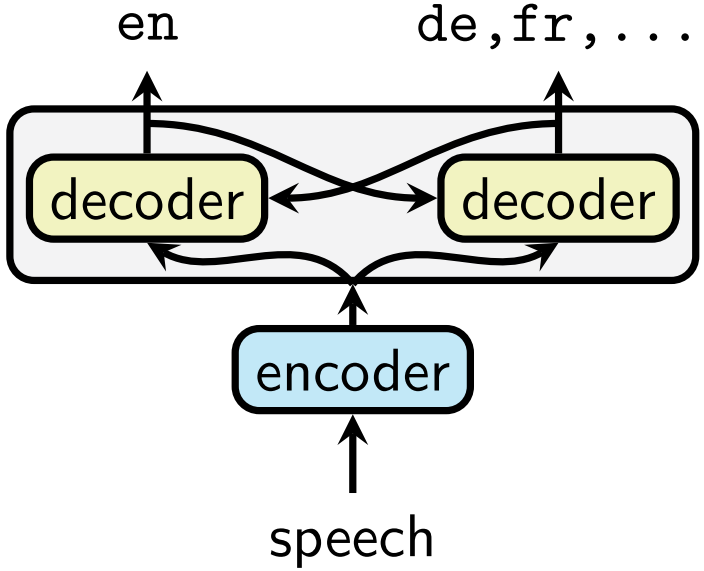
Special cases



Multitask

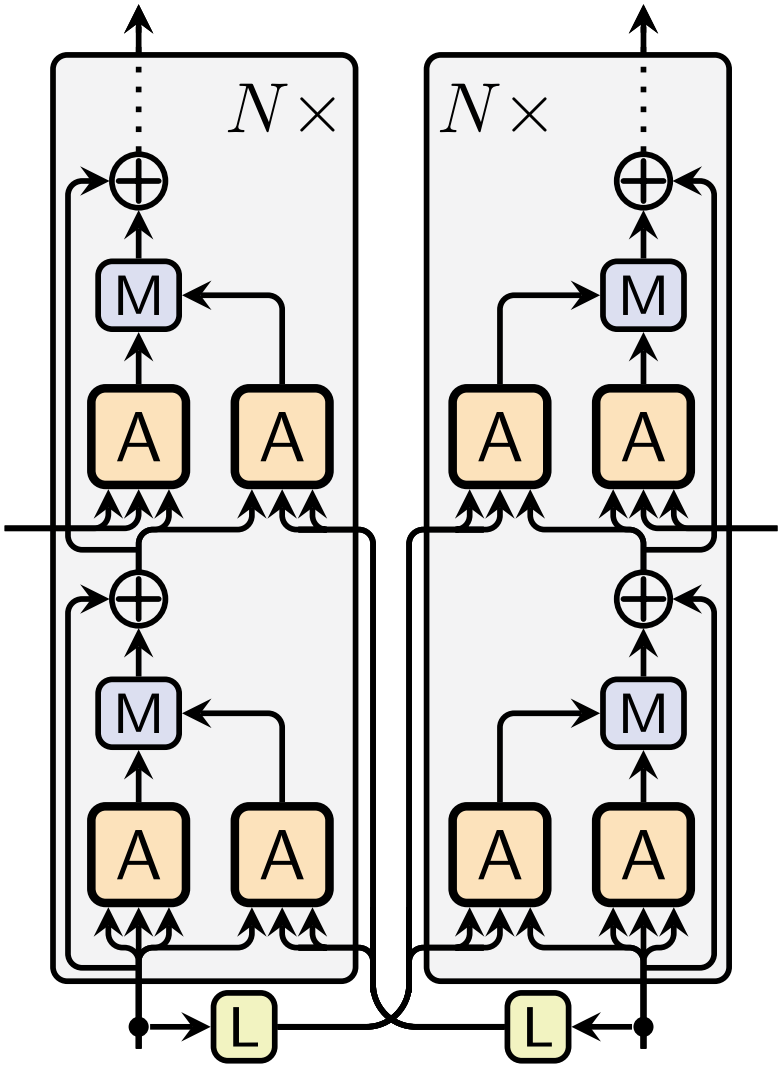


Two-stage/Triangle



Interactive

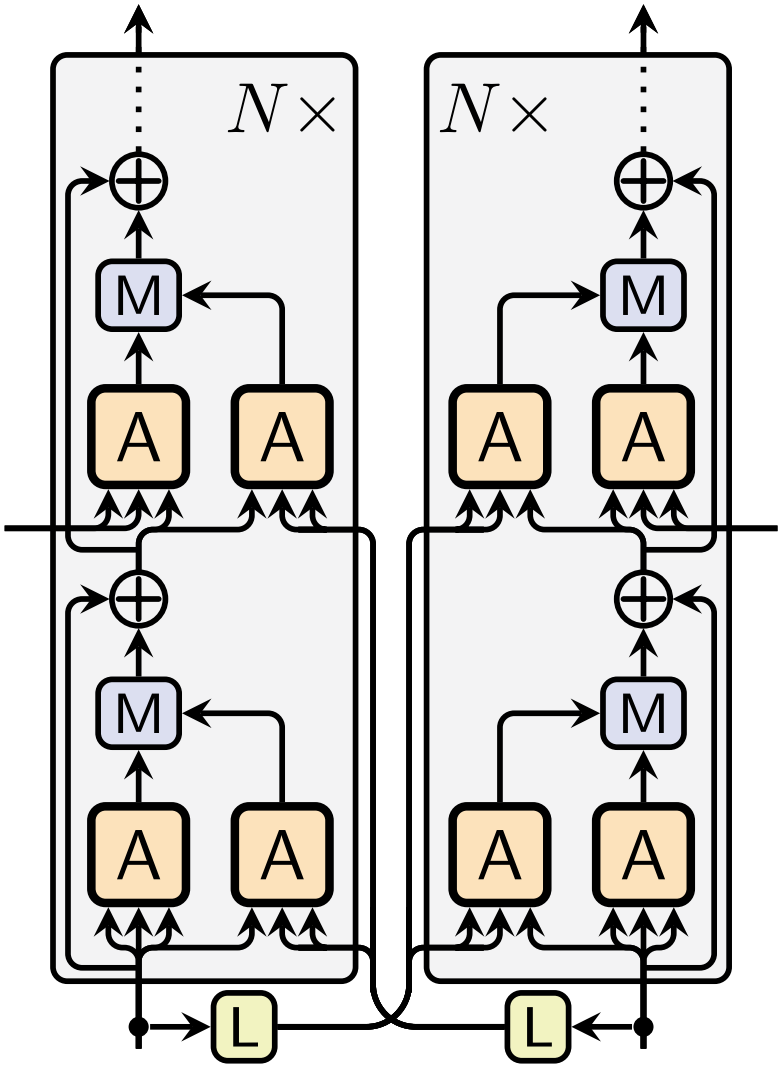
Cross dual-decoder Transformer



A (Attention), **M** (Merge), **L** (LayerNorm).

- Two common Transformer decoders.

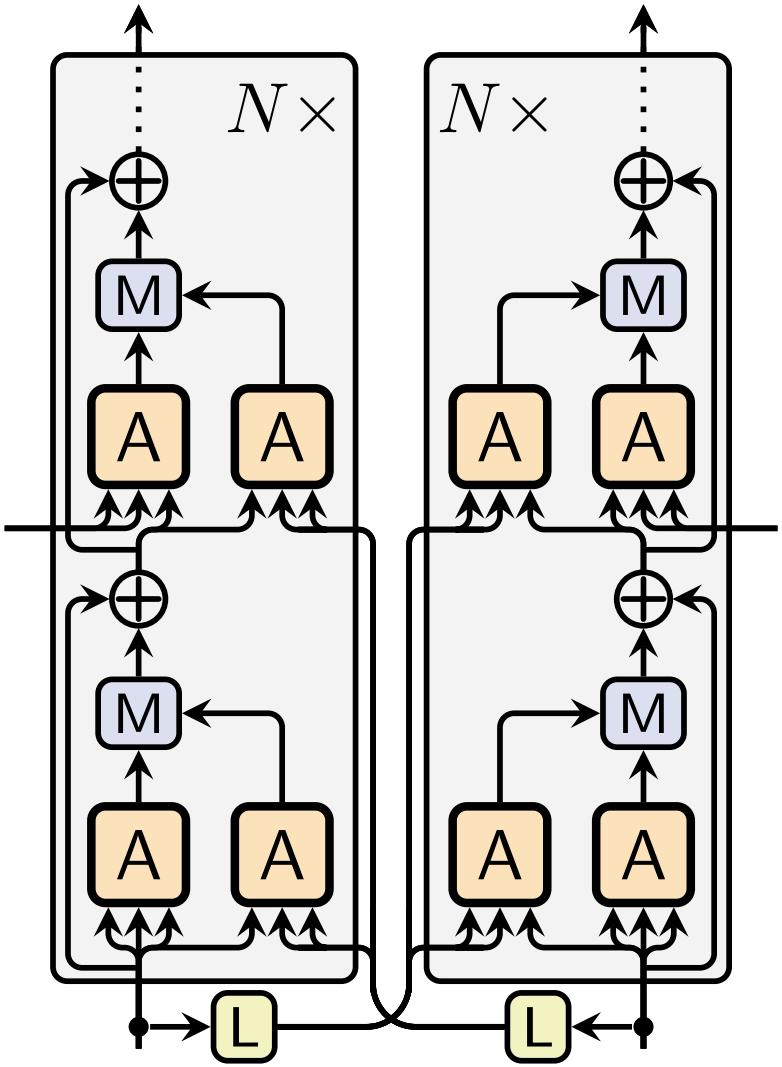
Cross dual-decoder Transformer



A (Attention), **M** (Merge), **L** (LayerNorm).

- Two common Transformer decoders.
- Four **dual-attention** layers:
  - two *dual-attention at self*; and
  - two *dual-attention at source*.

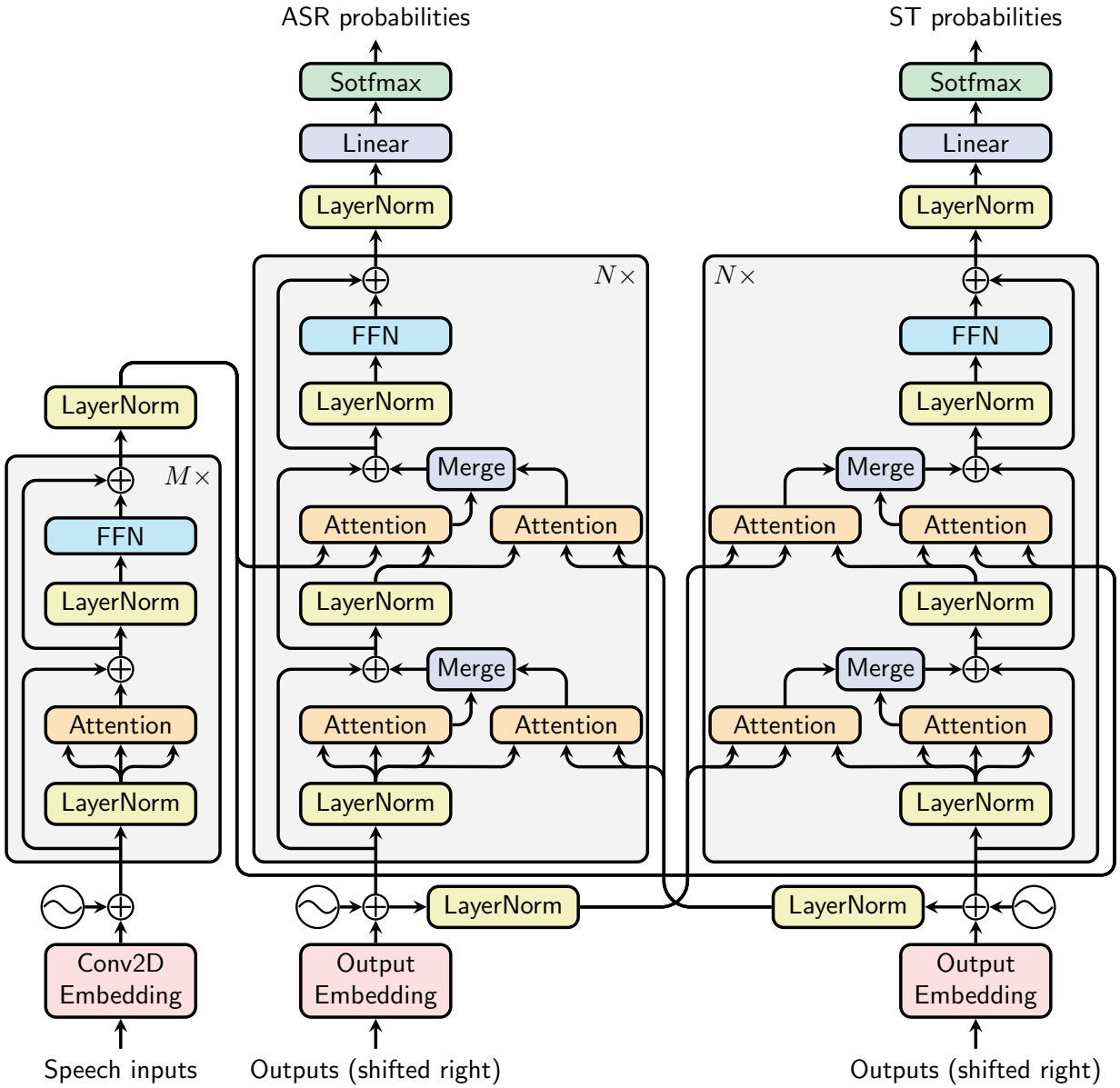
Cross dual-decoder Transformer



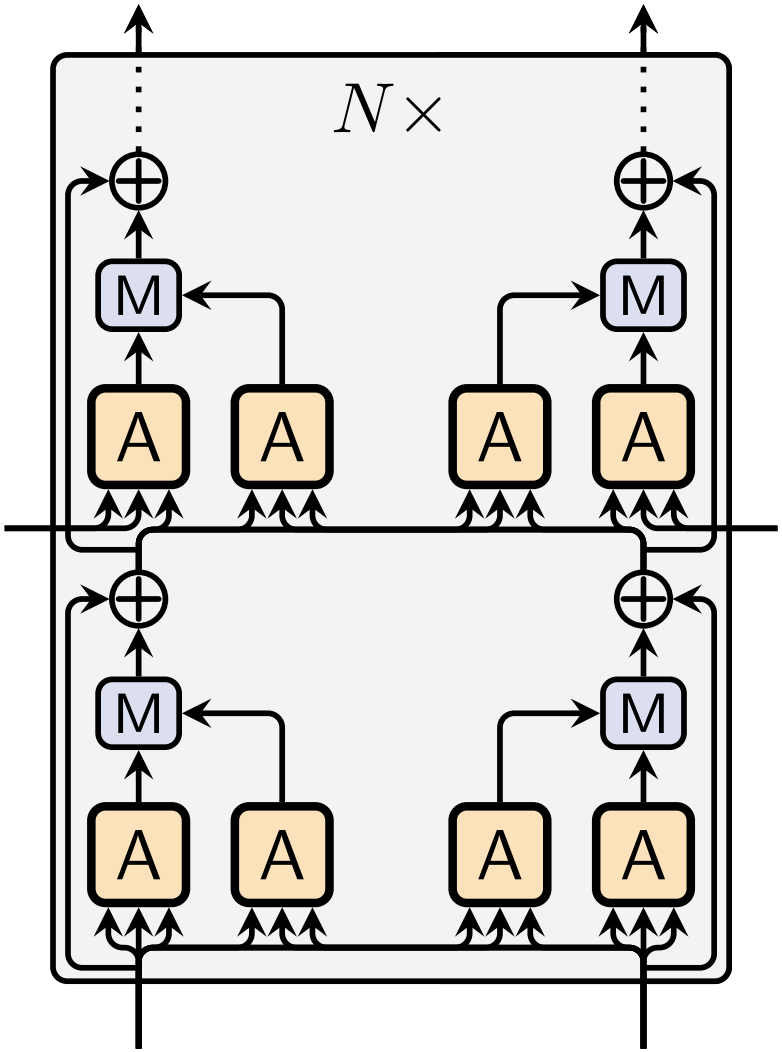
A (Attention), **M** (Merge), **L** (LayerNorm).

- Two common Transformer decoders.
- Four **dual-attention** layers:
  - two *dual-attention at self*; and
  - two *dual-attention at source*.
- Each dual-attention layer:
  - Query **Q**: from the main branch.
  - Key **K**, Value **V**: from the *previous decoding step outputs* of the other decoder.

Cross dual-decoder Transformer



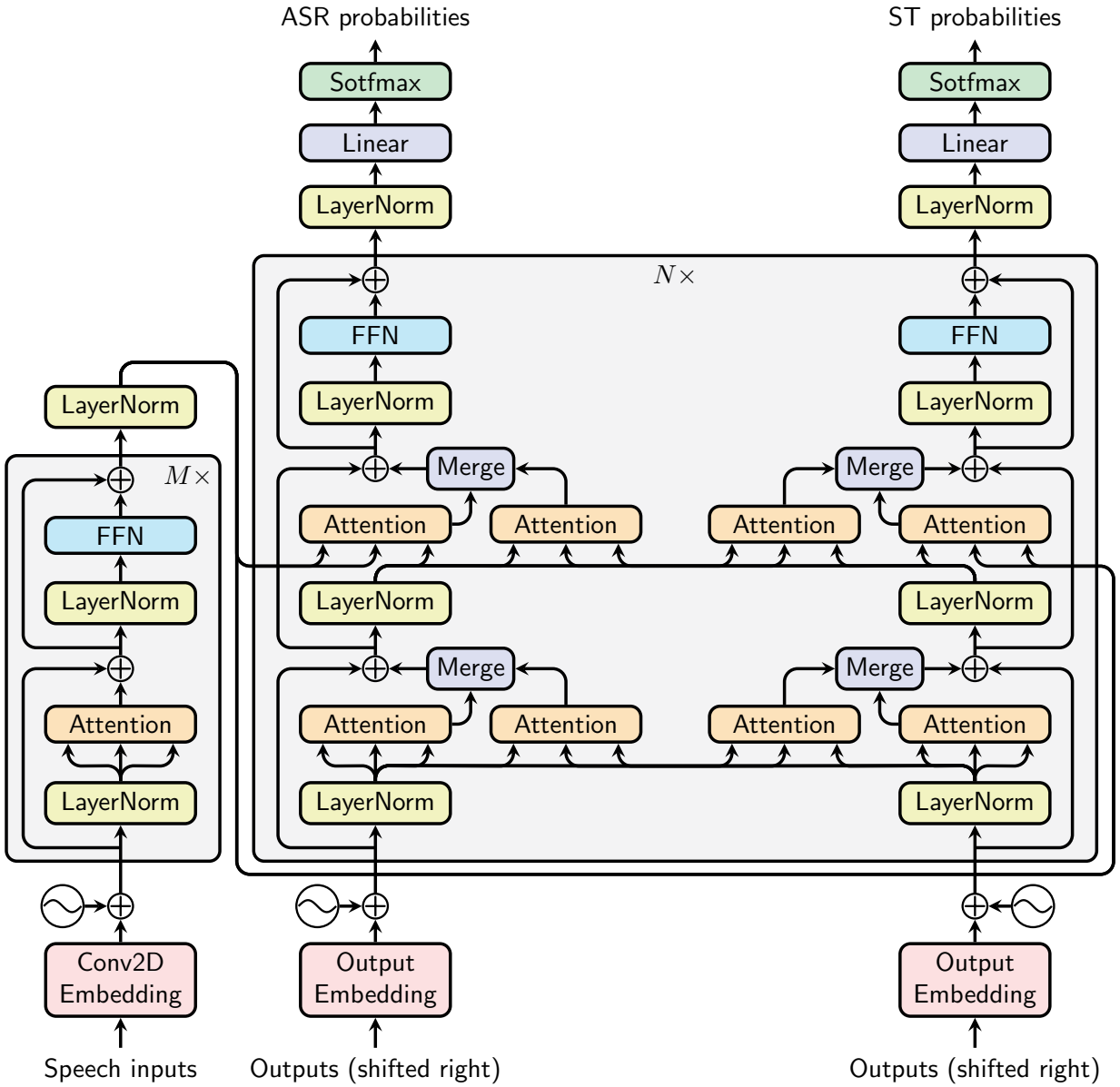
Parallel dual-decoder Transformer



A (Attention), **M** (Merge), **L** (LayerNorm).

- *Higher level of dependency* between the two decoders.
- Each dual-attention layer:
  - Query **Q**: from the main branch.
  - Key **K**, Value **V**: *from the other decoder at the same level*.

Parallel dual-decoder Transformer



Dual-decoder variants

Dual-decoder variants

1. **Asymmetric dual-decoder:** either ASR attends ST or the inverse, but not both.

Dual-decoder variants

1. **Asymmetric dual-decoder:** either ASR attends ST or the inverse, but not both.
2. **At-self or at-source dual-attention:** removing either the dual-attention at source or the dual-attention at self.

Dual-decoder variants

1. **Asymmetric dual-decoder:** either ASR attends ST or the inverse, but not both.
2. **At-self or at-source dual-attention:** removing either the dual-attention at source or the dual-attention at self.
3. **Merging operators:**

$$\mathbf{H}_{\text{out}} = \text{Merge}(\mathbf{H}_{\text{main}}, \mathbf{H}_{\text{dual}}) \triangleq \begin{cases} \mathbf{H}_{\text{main}} & \text{if no dual-attention,} \\ \mathbf{H}_{\text{main}} + \lambda \mathbf{H}_{\text{dual}}, & \text{if } \texttt{sum} \text{ operator,} \\ \text{linear}\left([\mathbf{H}_{\text{main}}; \mathbf{H}_{\text{dual}}]\right) & \text{if } \texttt{concat} \text{ operator.} \end{cases}$$

For the `sum` operator, in particular, we perform experiments for *learnable* or *fixed* λ.

Training and Decoding

- **Loss function:** $L(\hat{\mathbf{y}}, \hat{\mathbf{z}}, \mathbf{y}, \mathbf{z}) = \alpha L_{\text{asr}}(\hat{\mathbf{y}}, \mathbf{y}) + (1 - \alpha) L_{\text{st}}(\hat{\mathbf{z}}, \mathbf{z})$,
  - $(\mathbf{y}, \mathbf{z})$: ground-truths ($\mathbf{y}$ for ASR, $\mathbf{z}$ for ST).
  - $(\hat{\mathbf{y}}, \hat{\mathbf{z}})$: predictions.

Training and Decoding

- **Loss function:** $L(\hat{\mathbf{y}}, \hat{\mathbf{z}}, \mathbf{y}, \mathbf{z}) = \alpha L_{\text{asr}}(\hat{\mathbf{y}}, \mathbf{y}) + (1 - \alpha) L_{\text{st}}(\hat{\mathbf{z}}, \mathbf{z})$,
  - $(\mathbf{y}, \mathbf{z})$: ground-truths ($\mathbf{y}$ for ASR, $\mathbf{z}$ for ST).
  - $(\hat{\mathbf{y}}, \hat{\mathbf{z}})$: predictions.
- **Batching:** each mini-batch contains all languages.

Training and Decoding

- **Loss function:** $L(\hat{\mathbf{y}}, \hat{\mathbf{z}}, \mathbf{y}, \mathbf{z}) = \alpha L_{\text{asr}}(\hat{\mathbf{y}}, \mathbf{y}) + (1 - \alpha) L_{\text{st}}(\hat{\mathbf{z}}, \mathbf{z})$,
  - $(\mathbf{y}, \mathbf{z})$: ground-truths ($\mathbf{y}$ for ASR, $\mathbf{z}$ for ST).
  - $(\hat{\mathbf{y}}, \hat{\mathbf{z}})$: predictions.
- **Batching:** each mini-batch contains all languages.
- **Target forcing:** a *language-specific token* is prepended to the target sentence.

Training and Decoding

- **Loss function:** $L(\hat{\mathbf{y}}, \hat{\mathbf{z}}, \mathbf{y}, \mathbf{z}) = \alpha L_{\text{asr}}(\hat{\mathbf{y}}, \mathbf{y}) + (1 - \alpha) L_{\text{st}}(\hat{\mathbf{z}}, \mathbf{z})$,
  - $(\mathbf{y}, \mathbf{z})$: ground-truths ($\mathbf{y}$ for ASR, $\mathbf{z}$ for ST).
  - $(\hat{\mathbf{y}}, \hat{\mathbf{z}})$: predictions.
- **Batching:** each mini-batch contains all languages.
- **Target forcing:** a *language-specific token* is prepended to the target sentence.
- **Decoding:** *single joint beam*.
  - $\rightarrow$ Each hypothesis: tuple of ASR and ST sub-hypotheses.
  - $\rightarrow$ Score: sum of log probabilities of ASR, ST outputs.
  - $\rightarrow$ Best B hypotheses with highest score are retained.

Implementation details

- **Dataset:** MuST-C. English to 8 languages: Dutch, French, German, Italian, Portuguese, Romanian, Russian, and Spanish.

Implementation details

- **Dataset:** MuST-C. English to 8 languages: Dutch, French, German, Italian, Portuguese, Romanian, Russian, and Spanish.
- **Models:** 12-layer encoder + 6-layer decoders, except for `independent++` (8 layers).

Implementation details

- **Dataset:** MuST-C. English to 8 languages: Dutch, French, German, Italian, Portuguese, Romanian, Russian, and Spanish.
- **Models:** 12-layer encoder + 6-layer decoders, except for `independent++` (8 layers).
- **Text preprocessing:** Normalized and tokenized using Moses. Transcription lower-cased. Punctuation removed.

Implementation details

- **Dataset:** MuST-C. English to 8 languages: Dutch, French, German, Italian, Portuguese, Romanian, Russian, and Spanish.
- **Models:** 12-layer encoder + 6-layer decoders, except for `independent++` (8 layers).
- **Text preprocessing:** Normalized and tokenized using Moses. Transcription lower-cased. Punctuation removed.
- **Vocabulary:** *joint BPE 8K*.

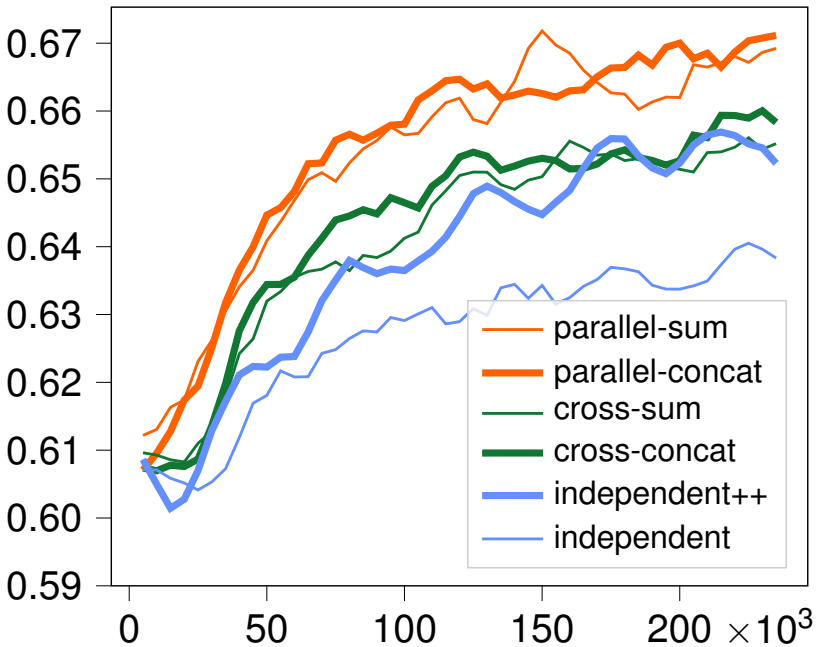
Implementation details

- **Dataset:** MuST-C. English to 8 languages: Dutch, French, German, Italian, Portuguese, Romanian, Russian, and Spanish.
- **Models:** 12-layer encoder + 6-layer decoders, except for `independent++` (8 layers).
- **Text preprocessing:** Normalized and tokenized using Moses. Transcription lower-cased. Punctuation removed.
- **Vocabulary:** *joint BPE 8K*.
- **Speech pre-processing:**
  - 80-*d* *log Mel filter-bank* coefficients + 3-*d* pitch features.
  - *Speed pertubation* with three factors of 0.9, 1.0, 1.1.
  - *SpecAugment* with $W = 5, T = 40, F = 30$.

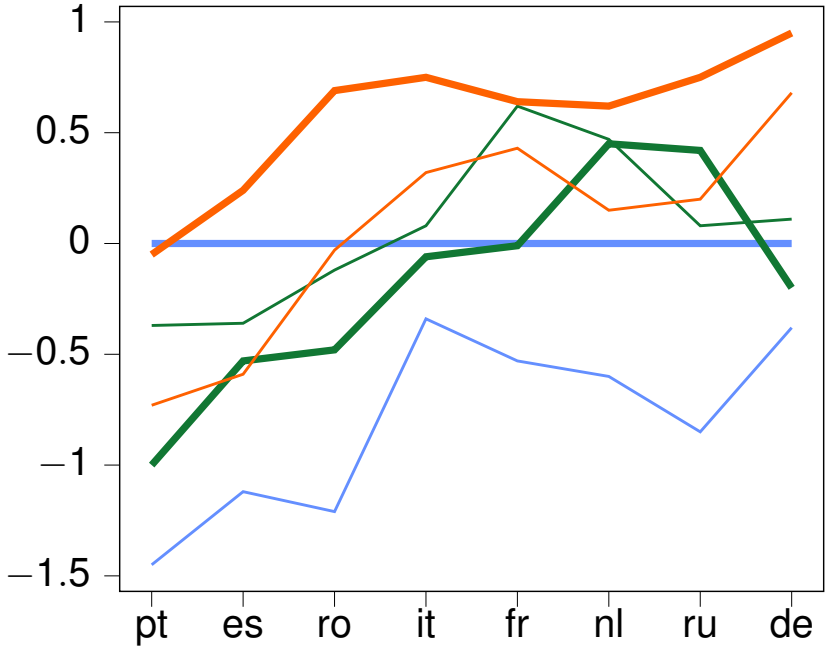
Implementation details

- **Dataset:** MuST-C. English to 8 languages: Dutch, French, German, Italian, Portuguese, Romanian, Russian, and Spanish.
- **Models:** 12-layer encoder + 6-layer decoders, except for `independent++` (8 layers).
- **Text preprocessing:** Normalized and tokenized using Moses. Transcription lower-cased. Punctuation removed.
- **Vocabulary:** *joint BPE 8K*.
- **Speech pre-processing:**
  - 80-*d* *log Mel filter-bank* coefficients + 3-*d* pitch features.
  - *Speed pertubation* with three factors of 0.9, 1.0, 1.1.
  - *SpecAugment* with $W = 5, T = 40, F = 30$.
- **Evaluation:** *BLEU/sacreBLEU* for translation, *WER* (word error rate) for recognition.

Validation accuracy and relative BLEU improvement

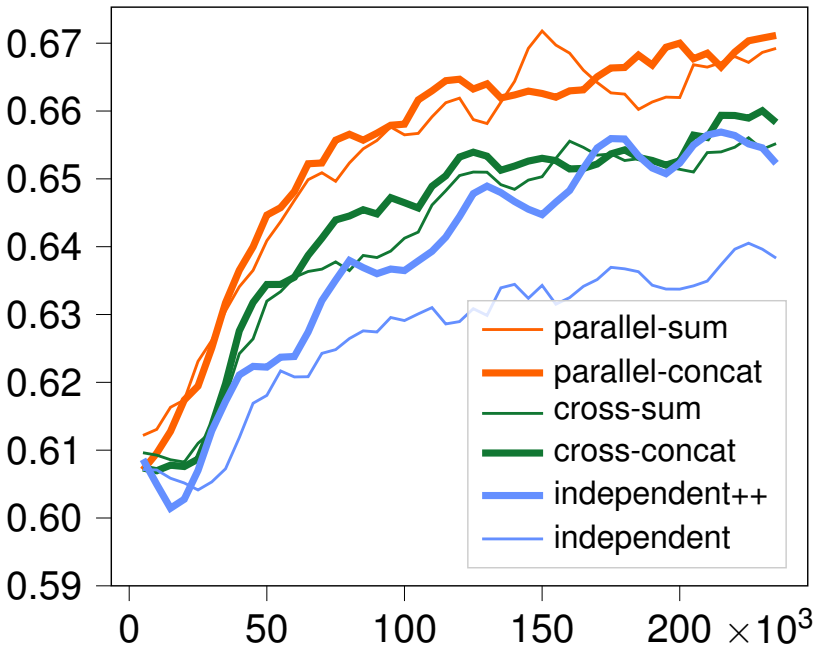


(a) ST validation accuracy per training step.

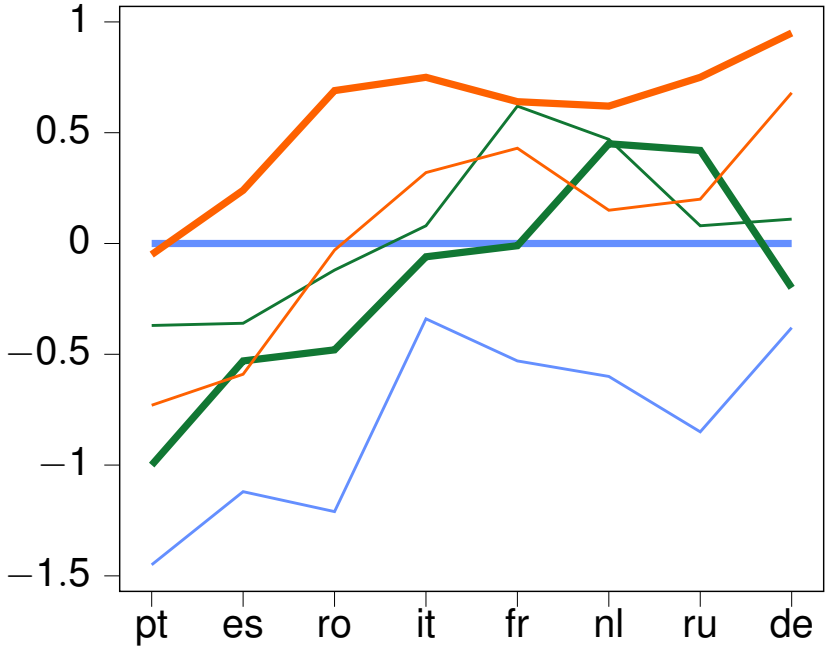


(b) Relative BLEU on MuST-C `dev` set.

Validation accuracy and relative BLEU improvement



(a) ST validation accuracy per training step.



(b) Relative BLEU on MuST-C `dev` set.

- Parallel models consistently outperform others in terms of validation accuracy.

Detailed results

- Blue means better than strongest baseline (`independent++`).
- `crx`: cross dual-decoder, `par`: parallel dual-decoder.

No	type	side	self	src	merge	params	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	independent (shared)					31.3M	19.40	27.77	24.65	19.93	21.53	24.24	18.19	10.99	20.84	14.2
2	independent					44.8M	20.11	28.18	25.61	20.76	21.83	25.45	18.45	11.31	21.46	12.6
3	independent++					51.2M	20.25	29.48	26.10	21.05	22.34	**26.71**	19.67	12.10	22.21	12.9

* no normalization for dual-attention input, † `sum` merging has $\lambda = 0.3$ fixed,
^{R3} ASR is 3 steps ahead of ST, ^{T3} ST is 3 steps ahead of ASR.

Detailed results

- Blue means better than strongest baseline (`independent++`).
- `crx`: cross dual-decoder, `par`: parallel dual-decoder.

No	type	side	self	src	merge	params	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	independent (shared)					31.3M	19.40	27.77	24.65	19.93	21.53	24.24	18.19	10.99	20.84	14.2
2	independent					44.8M	20.11	28.18	25.61	20.76	21.83	25.45	18.45	11.31	21.46	12.6
3	independent++					51.2M	20.25	29.48	26.10	21.05	22.34	**26.71**	19.67	12.10	22.21	12.9
4	crx	st	-	✓	sum	46.4M	20.01	28.57	25.86	20.66	22.26	25.36	19.06	12.00	21.72	12.7
5	crx	both	-	✓	concat	51.2M	20.36	28.51	25.80	21.18	22.10	25.24	19.55	11.89	21.83	12.3
6	crx	both	-	✓	sum	48.0M	19.99	28.87	26.09	20.94	21.67	25.42	18.85	11.83	21.71	12.2
7	crx	both	✓	✓	concat	54.3M	20.07	28.73	26.01	20.93	22.59	25.60	19.08	12.46	21.93	12.4
8	crx	both	✓	✓	sum	51.2M	20.38	28.90	26.64	21.07	22.61	26.23	19.44	12.12	22.17	**12.1**
9	crx*	both	✓	-	sum	48.0M	19.72	27.96	25.49	20.52	21.56	25.01	18.53	11.33	21.26	12.8
10	crx*	both	✓	-	sum†	48.0M	18.62	27.11	24.41	19.73	20.47	24.49	17.23	11.09	20.39	12.8
11	crx*	both	✓	✓	sum	51.2M	19.54	28.17	25.68	20.95	21.55	24.77	18.76	11.28	21.34	12.3

*no normalization for dual-attention input, †`sum` merging has $\lambda = 0.3$ fixed,
R3ASR is 3 steps ahead of ST, T3ST is 3 steps ahead of ASR.

Detailed results

- Blue means better than strongest baseline (`independent++`).
- `crx`: cross dual-decoder, `par`: parallel dual-decoder.

No	type	side	self	src	merge	params	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	independent (shared)					31.3M	19.40	27.77	24.65	19.93	21.53	24.24	18.19	10.99	20.84	14.2
2	independent					44.8M	20.11	28.18	25.61	20.76	21.83	25.45	18.45	11.31	21.46	12.6
3	independent++					51.2M	20.25	29.48	26.10	21.05	22.34	**26.71**	19.67	12.10	22.21	12.9
4	crx	st	-	✓	sum	46.4M	20.01	28.57	25.86	20.66	22.26	25.36	19.06	12.00	21.72	12.7
5	crx	both	-	✓	concat	51.2M	20.36	28.51	25.80	21.18	22.10	25.24	19.55	11.89	21.83	12.3
6	crx	both	-	✓	sum	48.0M	19.99	28.87	26.09	20.94	21.67	25.42	18.85	11.83	21.71	12.2
7	crx	both	✓	✓	concat	54.3M	20.07	28.73	26.01	20.93	22.59	25.60	19.08	12.46	21.93	12.4
8	crx	both	✓	✓	sum	51.2M	20.38	28.90	26.64	21.07	22.61	26.23	19.44	12.12	22.17	**12.1**
9	crx*	both	✓	-	sum	48.0M	19.72	27.96	25.49	20.52	21.56	25.01	18.53	11.33	21.26	12.8
10	crx*	both	✓	-	sum†	48.0M	18.62	27.11	24.41	19.73	20.47	24.49	17.23	11.09	20.39	12.8
11	crx*	both	✓	✓	sum	51.2M	19.54	28.17	25.68	20.95	21.55	24.77	18.76	11.28	21.34	12.3
12	par	st	✓	✓	concat	49.6M	20.57	28.84	26.08	20.85	22.11	25.70	19.36	11.90	21.93	13.0
13	par	both	-	✓	concat	51.2M	20.84	29.51	26.44	21.53	22.68	25.94	19.04	12.60	22.32	12.5
14	par	both	-	✓	sum	48.0M	20.85	29.18	26.38	**22.14**	22.87	26.49	19.70	12.74	22.54	12.7
15	par	both	✓	-	sum	48.0M	20.56	29.21	26.54	21.07	22.51	25.75	19.64	**12.80**	22.26	12.8
16	par	both	✓	✓	concat	54.3M	**21.22**	29.50	**26.66**	21.74	22.76	26.66	**20.25**	12.79	22.70	12.7
17	par	both	✓	✓	sum	51.2M	20.95	28.67	26.45	21.31	22.29	25.87	19.53	12.24	22.16	12.8
18	par	both^{R3}	-	✓	sum	48.0M	**21.22**	**30.12**	26.53	22.06	**23.37**	26.59	19.82	12.54	**22.78**	12.6
19	par	both^{T3}	-	✓	sum	48.0M	20.35	28.61	25.94	21.22	22.12	25.19	19.36	11.99	21.85	13.6

* no normalization for dual-attention input, † `sum` merging has $\lambda = 0.3$ fixed,
R3 ASR is 3 steps ahead of ST, T3 ST is 3 steps ahead of ASR.

Detailed results

Parallel dual-decoder is better than independent decoder in both BLEUs and WERs $\rightarrow$ no BLEU-WER trade-off.

No	type	side	self	src	merge	params	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	independent (shared)					31.3M	19.40	27.77	24.65	19.93	21.53	24.24	18.19	10.99	20.84	14.2
2	independent					44.8M	20.11	28.18	25.61	20.76	21.83	25.45	18.45	11.31	21.46	12.6
3	independent++					51.2M	20.25	29.48	26.10	21.05	22.34	**26.71**	19.67	12.10	22.21	12.9
4	crx	st	-	✓	sum	46.4M	20.01	28.57	25.86	20.66	22.26	25.36	19.06	12.00	21.72	12.7
5	crx	both	-	✓	concat	51.2M	20.36	28.51	25.80	21.18	22.10	25.24	19.55	11.89	21.83	12.3
6	crx	both	-	✓	sum	48.0M	19.99	28.87	26.09	20.94	21.67	25.42	18.85	11.83	21.71	12.2
7	crx	both	✓	✓	concat	54.3M	20.07	28.73	26.01	20.93	22.59	25.60	19.08	12.46	21.93	12.4
8	crx	both	✓	✓	sum	51.2M	20.38	28.90	26.64	21.07	22.61	26.23	19.44	12.12	22.17	**12.1**
9	crx*	both	✓	-	sum	48.0M	19.72	27.96	25.49	20.52	21.56	25.01	18.53	11.33	21.26	12.8
10	crx*	both	✓	-	sum†	48.0M	18.62	27.11	24.41	19.73	20.47	24.49	17.23	11.09	20.39	12.8
11	crx*	both	✓	✓	sum	51.2M	19.54	28.17	25.68	20.95	21.55	24.77	18.76	11.28	21.34	12.3
12	par	st	✓	✓	concat	49.6M	20.57	28.84	26.08	20.85	22.11	25.70	19.36	11.90	21.93	13.0
13	par	both	-	✓	concat	51.2M	20.84	29.51	26.44	21.53	22.68	25.94	19.04	12.60	22.32	12.5
14	par	both	-	✓	sum	48.0M	20.85	29.18	26.38	**22.14**	22.87	26.49	19.70	12.74	22.54	12.7
15	par	both	✓	-	sum	48.0M	20.56	29.21	26.54	21.07	22.51	25.75	19.64	**12.80**	22.26	12.8
16	par	both	✓	✓	concat	54.3M	**21.22**	29.50	**26.66**	21.74	22.76	26.66	**20.25**	12.79	22.70	12.7
17	par	both	✓	✓	sum	51.2M	20.95	28.67	26.45	21.31	22.29	25.87	19.53	12.24	22.16	12.8
18	par	both^{R3}	-	✓	sum	48.0M	**21.22**	**30.12**	26.53	22.06	**23.37**	26.59	19.82	12.54	**22.78**	12.6
19	par	both^{T3}	-	✓	sum	48.0M	20.35	28.61	25.94	21.22	22.12	25.19	19.36	11.99	21.85	13.6

* no normalization for dual-attention input, † `sum` merging has $\lambda = 0.3$ fixed,
R3 ASR is 3 steps ahead of ST, T3 ST is 3 steps ahead of ASR.

Detailed results

Cross dual-decoder is better than the weak baseline (`independent`), but not better than the strong baseline (`independent++`).

No	type	side	self	src	merge	params	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	independent (shared)					31.3M	19.40	27.77	24.65	19.93	21.53	24.24	18.19	10.99	20.84	14.2
2	independent					44.8M	20.11	28.18	25.61	20.76	21.83	25.45	18.45	11.31	21.46	12.6
3	independent++					51.2M	20.25	29.48	26.10	21.05	22.34	**26.71**	19.67	12.10	22.21	12.9
4	crx	st	-	✓	sum	46.4M	20.01	28.57	25.86	20.66	22.26	25.36	19.06	12.00	21.72	12.7
5	crx	both	-	✓	concat	51.2M	20.36	28.51	25.80	21.18	22.10	25.24	19.55	11.89	21.83	12.3
6	crx	both	-	✓	sum	48.0M	19.99	28.87	26.09	20.94	21.67	25.42	18.85	11.83	21.71	12.2
7	crx	both	✓	✓	concat	54.3M	20.07	28.73	26.01	20.93	22.59	25.60	19.08	12.46	21.93	12.4
8	crx	both	✓	✓	sum	51.2M	20.38	28.90	26.64	21.07	22.61	26.23	19.44	12.12	22.17	**12.1**
9	crx*	both	✓	-	sum	48.0M	19.72	27.96	25.49	20.52	21.56	25.01	18.53	11.33	21.26	12.8
10	crx*	both	✓	-	sum†	48.0M	18.62	27.11	24.41	19.73	20.47	24.49	17.23	11.09	20.39	12.8
11	crx*	both	✓	✓	sum	51.2M	19.54	28.17	25.68	20.95	21.55	24.77	18.76	11.28	21.34	12.3
12	par	st	✓	✓	concat	49.6M	20.57	28.84	26.08	20.85	22.11	25.70	19.36	11.90	21.93	13.0
13	par	both	-	✓	concat	51.2M	20.84	29.51	26.44	21.53	22.68	25.94	19.04	12.60	22.32	12.5
14	par	both	-	✓	sum	48.0M	20.85	29.18	26.38	**22.14**	22.87	26.49	19.70	12.74	22.54	12.7
15	par	both	✓	-	sum	48.0M	20.56	29.21	26.54	21.07	22.51	25.75	19.64	**12.80**	22.26	12.8
16	par	both	✓	✓	concat	54.3M	**21.22**	29.50	**26.66**	21.74	22.76	26.66	**20.25**	12.79	22.70	12.7
17	par	both	✓	✓	sum	51.2M	20.95	28.67	26.45	21.31	22.29	25.87	19.53	12.24	22.16	12.8
18	par	both^{R3}	-	✓	sum	48.0M	**21.22**	**30.12**	26.53	22.06	**23.37**	26.59	19.82	12.54	**22.78**	12.6
19	par	both^{T3}	-	✓	sum	48.0M	20.35	28.61	25.94	21.22	22.12	25.19	19.36	11.99	21.85	13.6

* no normalization for dual-attention input, † `sum` merging has $\lambda = 0.3$ fixed,
R3 ASR is 3 steps ahead of ST, T3 ST is 3 steps ahead of ASR.

Detailed results

Cross dual-decoder is better than the weak baseline (`independent`), but not better than the strong baseline (`independent++`).

No	type	side	self	src	merge	params	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	independent (shared)					31.3M	19.40	27.77	24.65	19.93	21.53	24.24	18.19	10.99	20.84	14.2
2	independent					44.8M	20.11	28.18	25.61	20.76	21.83	25.45	18.45	11.31	21.46	12.6
3	independent++					51.2M	20.25	29.48	26.10	21.05	22.34	**26.71**	19.67	12.10	22.21	12.9
4	crx	st	-	✓	sum	46.4M	20.01	28.57	25.86	20.66	22.26	25.36	19.06	12.00	21.72	12.7
5	crx	both	-	✓	concat	51.2M	20.36	28.51	25.80	21.18	22.10	25.24	19.55	11.89	21.83	12.3
6	crx	both	-	✓	sum	48.0M	19.99	28.87	26.09	20.94	21.67	25.42	18.85	11.83	21.71	12.2
7	crx	both	✓	✓	concat	54.3M	20.07	28.73	26.01	20.93	22.59	25.60	19.08	12.46	21.93	12.4
8	crx	both	✓	✓	sum	51.2M	20.38	28.90	26.64	21.07	22.61	26.23	19.44	12.12	22.17	**12.1**
9	crx*	both	✓	-	sum	48.0M	19.72	27.96	25.49	20.52	21.56	25.01	18.53	11.33	21.26	12.8
10	crx*	both	✓	-	sum†	48.0M	18.62	27.11	24.41	19.73	20.47	24.49	17.23	11.09	20.39	12.8
11	crx*	both	✓	✓	sum	51.2M	19.54	28.17	25.68	20.95	21.55	24.77	18.76	11.28	21.34	12.3
12	par	st	✓	✓	concat	49.6M	20.57	28.84	26.08	20.85	22.11	25.70	19.36	11.90	21.93	13.0
13	par	both	-	✓	concat	51.2M	20.84	29.51	26.44	21.53	22.68	25.94	19.04	12.60	22.32	12.5
14	par	both	-	✓	sum	48.0M	20.85	29.18	26.38	**22.14**	22.87	26.49	19.70	12.74	22.54	12.7
15	par	both	✓	-	sum	48.0M	20.56	29.21	26.54	21.07	22.51	25.75	19.64	**12.80**	22.26	12.8
16	par	both	✓	✓	concat	54.3M	**21.22**	29.50	**26.66**	21.74	22.76	26.66	**20.25**	12.79	22.70	12.7
17	par	both	✓	✓	sum	51.2M	20.95	28.67	26.45	21.31	22.29	25.87	19.53	12.24	22.16	12.8
18	par	both^{R3}	-	✓	sum	48.0M	**21.22**	**30.12**	26.53	22.06	**23.37**	26.59	19.82	12.54	**22.78**	12.6
19	par	both^{T3}	-	✓	sum	48.0M	20.35	28.61	25.94	21.22	22.12	25.19	19.36	11.99	21.85	13.6

* no normalization for dual-attention input, † `sum` merging has $\lambda = 0.3$ fixed,
R3 ASR is 3 steps ahead of ST, T3 ST is 3 steps ahead of ASR.

Detailed results

Parallel model is best in terms of BLEUs.

No	type	side	self	src	merge	params	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	independent (shared)					31.3M	19.40	27.77	24.65	19.93	21.53	24.24	18.19	10.99	20.84	14.2
2	independent					44.8M	20.11	28.18	25.61	20.76	21.83	25.45	18.45	11.31	21.46	12.6
3	independent++					51.2M	20.25	29.48	26.10	21.05	22.34	**26.71**	19.67	12.10	22.21	12.9
4	crx	st	-	✓	sum	46.4M	20.01	28.57	25.86	20.66	22.26	25.36	19.06	12.00	21.72	12.7
5	crx	both	-	✓	concat	51.2M	20.36	28.51	25.80	21.18	22.10	25.24	19.55	11.89	21.83	12.3
6	crx	both	-	✓	sum	48.0M	19.99	28.87	26.09	20.94	21.67	25.42	18.85	11.83	21.71	12.2
7	crx	both	✓	✓	concat	54.3M	20.07	28.73	26.01	20.93	22.59	25.60	19.08	12.46	21.93	12.4
8	crx	both	✓	✓	sum	51.2M	20.38	28.90	26.64	21.07	22.61	26.23	19.44	12.12	22.17	**12.1**
9	crx*	both	✓	-	sum	48.0M	19.72	27.96	25.49	20.52	21.56	25.01	18.53	11.33	21.26	12.8
10	crx*	both	✓	-	sum†	48.0M	18.62	27.11	24.41	19.73	20.47	24.49	17.23	11.09	20.39	12.8
11	crx*	both	✓	✓	sum	51.2M	19.54	28.17	25.68	20.95	21.55	24.77	18.76	11.28	21.34	12.3
12	par	st	✓	✓	concat	49.6M	20.57	28.84	26.08	20.85	22.11	25.70	19.36	11.90	21.93	13.0
13	par	both	-	✓	concat	51.2M	20.84	29.51	26.44	21.53	22.68	25.94	19.04	12.60	22.32	12.5
14	par	both	-	✓	sum	48.0M	20.85	29.18	26.38	**22.14**	22.87	26.49	19.70	12.74	22.54	12.7
15	par	both	✓	-	sum	48.0M	20.56	29.21	26.54	21.07	22.51	25.75	19.64	**12.80**	22.26	12.8
16	par	both	✓	✓	concat	54.3M	**21.22**	29.50	**26.66**	21.74	22.76	26.66	**20.25**	12.79	22.70	12.7
17	par	both	✓	✓	sum	51.2M	20.95	28.67	26.45	21.31	22.29	25.87	19.53	12.24	22.16	12.8
18	par	both^{R3}	-	✓	sum	48.0M	**21.22**	**30.12**	26.53	22.06	**23.37**	26.59	19.82	12.54	**22.78**	12.6
19	par	both^{T3}	-	✓	sum	48.0M	20.35	28.61	25.94	21.22	22.12	25.19	19.36	11.99	21.85	13.6

* no normalization for dual-attention input, † $\texttt{sum}$ merging has $\lambda = 0.3$ fixed,
R3 ASR is 3 steps ahead of ST, T3 ST is 3 steps ahead of ASR.

Detailed results

Cross model is best in terms of WERs.

No	type	side	self	src	merge	params	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	independent (shared)					31.3M	19.40	27.77	24.65	19.93	21.53	24.24	18.19	10.99	20.84	14.2
2	independent					44.8M	20.11	28.18	25.61	20.76	21.83	25.45	18.45	11.31	21.46	12.6
3	independent++					51.2M	20.25	29.48	26.10	21.05	22.34	**26.71**	19.67	12.10	22.21	12.9
4	crx	st	-	✓	sum	46.4M	20.01	28.57	25.86	20.66	22.26	25.36	19.06	12.00	21.72	12.7
5	crx	both	-	✓	concat	51.2M	20.36	28.51	25.80	21.18	22.10	25.24	19.55	11.89	21.83	12.3
6	crx	both	-	✓	sum	48.0M	19.99	28.87	26.09	20.94	21.67	25.42	18.85	11.83	21.71	12.2
7	crx	both	✓	✓	concat	54.3M	20.07	28.73	26.01	20.93	22.59	25.60	19.08	12.46	21.93	12.4
8	crx	both	✓	✓	sum	51.2M	20.38	28.90	26.64	21.07	22.61	26.23	19.44	12.12	22.17	**12.1**
9	crx*	both	✓	-	sum	48.0M	19.72	27.96	25.49	20.52	21.56	25.01	18.53	11.33	21.26	12.8
10	crx*	both	✓	-	sum†	48.0M	18.62	27.11	24.41	19.73	20.47	24.49	17.23	11.09	20.39	12.8
11	crx*	both	✓	✓	sum	51.2M	19.54	28.17	25.68	20.95	21.55	24.77	18.76	11.28	21.34	12.3
12	par	st	✓	✓	concat	49.6M	20.57	28.84	26.08	20.85	22.11	25.70	19.36	11.90	21.93	13.0
13	par	both	-	✓	concat	51.2M	20.84	29.51	26.44	21.53	22.68	25.94	19.04	12.60	22.32	12.5
14	par	both	-	✓	sum	48.0M	20.85	29.18	26.38	**22.14**	22.87	26.49	19.70	12.74	22.54	12.7
15	par	both	✓	-	sum	48.0M	20.56	29.21	26.54	21.07	22.51	25.75	19.64	**12.80**	22.26	12.8
16	par	both	✓	✓	concat	54.3M	**21.22**	29.50	**26.66**	21.74	22.76	26.66	**20.25**	12.79	22.70	12.7
17	par	both	✓	✓	sum	51.2M	20.95	28.67	26.45	21.31	22.29	25.87	19.53	12.24	22.16	12.8
18	par	both^{R3}	-	✓	sum	48.0M	**21.22**	**30.12**	26.53	22.06	**23.37**	26.59	19.82	12.54	**22.78**	12.6
19	par	both^{T3}	-	✓	sum	48.0M	20.35	28.61	25.94	21.22	22.12	25.19	19.36	11.99	21.85	13.6

* no normalization for dual-attention input, † $\texttt{sum}$ merging has $\lambda = 0.3$ fixed,
R3 ASR is 3 steps ahead of ST, T3 ST is 3 steps ahead of ASR.

Detailed results

Symmetric better than asymmetric.

No	type	side	self	src	merge	params	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	independent (shared)					31.3M	19.40	27.77	24.65	19.93	21.53	24.24	18.19	10.99	20.84	14.2
2	independent					44.8M	20.11	28.18	25.61	20.76	21.83	25.45	18.45	11.31	21.46	12.6
3	independent++					51.2M	20.25	29.48	26.10	21.05	22.34	**26.71**	19.67	12.10	22.21	12.9
4	crx	st	-	✓	sum	46.4M	20.01	28.57	25.86	20.66	22.26	25.36	19.06	12.00	21.72	12.7
5	crx	both	-	✓	concat	51.2M	20.36	28.51	25.80	21.18	22.10	25.24	19.55	11.89	21.83	12.3
6	crx	both	-	✓	sum	48.0M	19.99	28.87	26.09	20.94	21.67	25.42	18.85	11.83	21.71	12.2
7	crx	both	✓	✓	concat	54.3M	20.07	28.73	26.01	20.93	22.59	25.60	19.08	12.46	21.93	12.4
8	crx	both	✓	✓	sum	51.2M	20.38	28.90	26.64	21.07	22.61	26.23	19.44	12.12	22.17	**12.1**
9	crx*	both	✓	-	sum	48.0M	19.72	27.96	25.49	20.52	21.56	25.01	18.53	11.33	21.26	12.8
10	crx*	both	✓	-	sum†	48.0M	18.62	27.11	24.41	19.73	20.47	24.49	17.23	11.09	20.39	12.8
11	crx*	both	✓	✓	sum	51.2M	19.54	28.17	25.68	20.95	21.55	24.77	18.76	11.28	21.34	12.3
12	par	st	✓	✓	concat	49.6M	20.57	28.84	26.08	20.85	22.11	25.70	19.36	11.90	21.93	13.0
13	par	both	-	✓	concat	51.2M	20.84	29.51	26.44	21.53	22.68	25.94	19.04	12.60	22.32	12.5
14	par	both	-	✓	sum	48.0M	20.85	29.18	26.38	**22.14**	22.87	26.49	19.70	12.74	22.54	12.7
15	par	both	✓	-	sum	48.0M	20.56	29.21	26.54	21.07	22.51	25.75	19.64	**12.80**	22.26	12.8
16	par	both	✓	✓	concat	54.3M	**21.22**	29.50	**26.66**	21.74	22.76	26.66	**20.25**	12.79	22.70	12.7
17	par	both	✓	✓	sum	51.2M	20.95	28.67	26.45	21.31	22.29	25.87	19.53	12.24	22.16	12.8
18	par	both^{R3}	-	✓	sum	48.0M	**21.22**	**30.12**	26.53	22.06	**23.37**	26.59	19.82	12.54	**22.78**	12.6
19	par	both^{T3}	-	✓	sum	48.0M	20.35	28.61	25.94	21.22	22.12	25.19	19.36	11.99	21.85	13.6

* no normalization for dual-attention input, † `sum` merging has $\lambda = 0.3$ fixed,
R3 ASR is 3 steps ahead of ST, T3 ST is 3 steps ahead of ASR.

Detailed results

Symmetric better than asymmetric.

No	type	side	self	src	merge	params	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	independent (shared)					31.3M	19.40	27.77	24.65	19.93	21.53	24.24	18.19	10.99	20.84	14.2
2	independent					44.8M	20.11	28.18	25.61	20.76	21.83	25.45	18.45	11.31	21.46	12.6
3	independent++					51.2M	20.25	29.48	26.10	21.05	22.34	**26.71**	19.67	12.10	22.21	12.9
4	crx	st	-	✓	sum	46.4M	20.01	28.57	25.86	20.66	22.26	25.36	19.06	12.00	21.72	12.7
5	crx	both	-	✓	concat	51.2M	20.36	28.51	25.80	21.18	22.10	25.24	19.55	11.89	21.83	12.3
6	crx	both	-	✓	sum	48.0M	19.99	28.87	26.09	20.94	21.67	25.42	18.85	11.83	21.71	12.2
7	crx	both	✓	✓	concat	54.3M	20.07	28.73	26.01	20.93	22.59	25.60	19.08	12.46	21.93	12.4
8	crx	both	✓	✓	sum	51.2M	20.38	28.90	26.64	21.07	22.61	26.23	19.44	12.12	22.17	**12.1**
9	crx*	both	✓	-	sum	48.0M	19.72	27.96	25.49	20.52	21.56	25.01	18.53	11.33	21.26	12.8
10	crx*	both	✓	-	sum†	48.0M	18.62	27.11	24.41	19.73	20.47	24.49	17.23	11.09	20.39	12.8
11	crx*	both	✓	✓	sum	51.2M	19.54	28.17	25.68	20.95	21.55	24.77	18.76	11.28	21.34	12.3
12	par	st	✓	✓	concat	49.6M	20.57	28.84	26.08	20.85	22.11	25.70	19.36	11.90	21.93	13.0
13	par	both	-	✓	concat	51.2M	20.84	29.51	26.44	21.53	22.68	25.94	19.04	12.60	22.32	12.5
14	par	both	-	✓	sum	48.0M	20.85	29.18	26.38	**22.14**	22.87	26.49	19.70	12.74	22.54	12.7
15	par	both	✓	-	sum	48.0M	20.56	29.21	26.54	21.07	22.51	25.75	19.64	**12.80**	22.26	12.8
16	par	both	✓	✓	concat	54.3M	**21.22**	29.50	**26.66**	21.74	22.76	26.66	**20.25**	12.79	22.70	12.7
17	par	both	✓	✓	sum	51.2M	20.95	28.67	26.45	21.31	22.29	25.87	19.53	12.24	22.16	12.8
18	par	both^{R3}	-	✓	sum	48.0M	**21.22**	**30.12**	26.53	22.06	**23.37**	26.59	19.82	12.54	**22.78**	12.6
19	par	both^{T3}	-	✓	sum	48.0M	20.35	28.61	25.94	21.22	22.12	25.19	19.36	11.99	21.85	13.6

* no normalization for dual-attention input, † `sum` merging has $\lambda = 0.3$ fixed,
R3 ASR is 3 steps ahead of ST, T3 ST is 3 steps ahead of ASR.

Detailed results

Learnable λ better than fixed λ.

No	type	side	self	src	merge	params	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	independent (shared)					31.3M	19.40	27.77	24.65	19.93	21.53	24.24	18.19	10.99	20.84	14.2
2	independent					44.8M	20.11	28.18	25.61	20.76	21.83	25.45	18.45	11.31	21.46	12.6
3	independent++					51.2M	20.25	29.48	26.10	21.05	22.34	**26.71**	19.67	12.10	22.21	12.9
4	crx	st	-	✓	sum	46.4M	20.01	28.57	25.86	20.66	22.26	25.36	19.06	12.00	21.72	12.7
5	crx	both	-	✓	concat	51.2M	20.36	28.51	25.80	21.18	22.10	25.24	19.55	11.89	21.83	12.3
6	crx	both	-	✓	sum	48.0M	19.99	28.87	26.09	20.94	21.67	25.42	18.85	11.83	21.71	12.2
7	crx	both	✓	✓	concat	54.3M	20.07	28.73	26.01	20.93	22.59	25.60	19.08	12.46	21.93	12.4
8	crx	both	✓	✓	sum	51.2M	20.38	28.90	26.64	21.07	22.61	26.23	19.44	12.12	22.17	**12.1**
9	crx*	both	✓	-	sum	48.0M	19.72	27.96	25.49	20.52	21.56	25.01	18.53	11.33	21.26	12.8
10	crx*	both	✓	-	sum†	48.0M	18.62	27.11	24.41	19.73	20.47	24.49	17.23	11.09	20.39	12.8
11	crx*	both	✓	✓	sum	51.2M	19.54	28.17	25.68	20.95	21.55	24.77	18.76	11.28	21.34	12.3
12	par	st	✓	✓	concat	49.6M	20.57	28.84	26.08	20.85	22.11	25.70	19.36	11.90	21.93	13.0
13	par	both	-	✓	concat	51.2M	20.84	29.51	26.44	21.53	22.68	25.94	19.04	12.60	22.32	12.5
14	par	both	-	✓	sum	48.0M	20.85	29.18	26.38	**22.14**	22.87	26.49	19.70	12.74	22.54	12.7
15	par	both	✓	-	sum	48.0M	20.56	29.21	26.54	21.07	22.51	25.75	19.64	**12.80**	22.26	12.8
16	par	both	✓	✓	concat	54.3M	**21.22**	29.50	**26.66**	21.74	22.76	26.66	**20.25**	12.79	22.70	12.7
17	par	both	✓	✓	sum	51.2M	20.95	28.67	26.45	21.31	22.29	25.87	19.53	12.24	22.16	12.8
18	par	both^{R3}	-	✓	sum	48.0M	**21.22**	**30.12**	26.53	22.06	**23.37**	26.59	19.82	12.54	**22.78**	12.6
19	par	both^{T3}	-	✓	sum	48.0M	20.35	28.61	25.94	21.22	22.12	25.19	19.36	11.99	21.85	13.6

* no normalization for dual-attention input, † `sum` merging has $\lambda = 0.3$ fixed,
R3 ASR is 3 steps ahead of ST, T3 ST is 3 steps ahead of ASR.

Detailed results

Wait-k: *ST-waits-for-ASR improves performance*, *ASR-waits-for-ST worsen*.

No	type	side	self	src	merge	params	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	independent (shared)					31.3M	19.40	27.77	24.65	19.93	21.53	24.24	18.19	10.99	20.84	14.2
2	independent					44.8M	20.11	28.18	25.61	20.76	21.83	25.45	18.45	11.31	21.46	12.6
3	independent++					51.2M	20.25	29.48	26.10	21.05	22.34	**26.71**	19.67	12.10	22.21	12.9
4	crx	st	-	✓	sum	46.4M	20.01	28.57	25.86	20.66	22.26	25.36	19.06	12.00	21.72	12.7
5	crx	both	-	✓	concat	51.2M	20.36	28.51	25.80	21.18	22.10	25.24	19.55	11.89	21.83	12.3
6	crx	both	-	✓	sum	48.0M	19.99	28.87	26.09	20.94	21.67	25.42	18.85	11.83	21.71	12.2
7	crx	both	✓	✓	concat	54.3M	20.07	28.73	26.01	20.93	22.59	25.60	19.08	12.46	21.93	12.4
8	crx	both	✓	✓	sum	51.2M	20.38	28.90	26.64	21.07	22.61	26.23	19.44	12.12	22.17	**12.1**
9	crx*	both	✓	-	sum	48.0M	19.72	27.96	25.49	20.52	21.56	25.01	18.53	11.33	21.26	12.8
10	crx*	both	✓	-	sum†	48.0M	18.62	27.11	24.41	19.73	20.47	24.49	17.23	11.09	20.39	12.8
11	crx*	both	✓	✓	sum	51.2M	19.54	28.17	25.68	20.95	21.55	24.77	18.76	11.28	21.34	12.3
12	par	st	✓	✓	concat	49.6M	20.57	28.84	26.08	20.85	22.11	25.70	19.36	11.90	21.93	13.0
13	par	both	-	✓	concat	51.2M	20.84	29.51	26.44	21.53	22.68	25.94	19.04	12.60	22.32	12.5
14	par	both	-	✓	sum	48.0M	20.85	29.18	26.38	**22.14**	22.87	26.49	19.70	12.74	22.54	12.7
15	par	both	✓	-	sum	48.0M	20.56	29.21	26.54	21.07	22.51	25.75	19.64	**12.80**	22.26	12.8
16	par	both	✓	✓	concat	54.3M	**21.22**	29.50	**26.66**	21.74	22.76	26.66	**20.25**	12.79	22.70	12.7
17	par	both	✓	✓	sum	51.2M	20.95	28.67	26.45	21.31	22.29	25.87	19.53	12.24	22.16	12.8
18	par	both^{R3}	-	✓	sum	48.0M	**21.22**	**30.12**	26.53	22.06	**23.37**	26.59	19.82	12.54	**22.78**	12.6
19	par	both^{T3}	-	✓	sum	48.0M	20.35	28.61	25.94	21.22	22.12	25.19	19.36	11.99	21.85	13.6

* no normalization for dual-attention input, † `sum` merging has $\lambda = 0.3$ fixed,
R3 ASR is 3 steps ahead of ST, T3 ST is 3 steps ahead of ASR.

Detailed results

Wait-k: *ST-waits-for-ASR improves performance*, *ASR-waits-for-ST worsen*.

No	type	side	self	src	merge	params	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	independent (shared)					31.3M	19.40	27.77	24.65	19.93	21.53	24.24	18.19	10.99	20.84	14.2
2	independent					44.8M	20.11	28.18	25.61	20.76	21.83	25.45	18.45	11.31	21.46	12.6
3	independent++					51.2M	20.25	29.48	26.10	21.05	22.34	**26.71**	19.67	12.10	22.21	12.9
4	crx	st	-	✓	sum	46.4M	20.01	28.57	25.86	20.66	22.26	25.36	19.06	12.00	21.72	12.7
5	crx	both	-	✓	concat	51.2M	20.36	28.51	25.80	21.18	22.10	25.24	19.55	11.89	21.83	12.3
6	crx	both	-	✓	sum	48.0M	19.99	28.87	26.09	20.94	21.67	25.42	18.85	11.83	21.71	12.2
7	crx	both	✓	✓	concat	54.3M	20.07	28.73	26.01	20.93	22.59	25.60	19.08	12.46	21.93	12.4
8	crx	both	✓	✓	sum	51.2M	20.38	28.90	26.64	21.07	22.61	26.23	19.44	12.12	22.17	**12.1**
9	crx*	both	✓	-	sum	48.0M	19.72	27.96	25.49	20.52	21.56	25.01	18.53	11.33	21.26	12.8
10	crx*	both	✓	-	sum†	48.0M	18.62	27.11	24.41	19.73	20.47	24.49	17.23	11.09	20.39	12.8
11	crx*	both	✓	✓	sum	51.2M	19.54	28.17	25.68	20.95	21.55	24.77	18.76	11.28	21.34	12.3
12	par	st	✓	✓	concat	49.6M	20.57	28.84	26.08	20.85	22.11	25.70	19.36	11.90	21.93	13.0
13	par	both	-	✓	concat	51.2M	20.84	29.51	26.44	21.53	22.68	25.94	19.04	12.60	22.32	12.5
14	par	both	-	✓	sum	48.0M	20.85	29.18	26.38	**22.14**	22.87	26.49	19.70	12.74	22.54	12.7
15	par	both	✓	-	sum	48.0M	20.56	29.21	26.54	21.07	22.51	25.75	19.64	**12.80**	22.26	12.8
16	par	both	✓	✓	concat	54.3M	**21.22**	29.50	**26.66**	21.74	22.76	26.66	**20.25**	12.79	22.70	12.7
17	par	both	✓	✓	sum	51.2M	20.95	28.67	26.45	21.31	22.29	25.87	19.53	12.24	22.16	12.8
18	par	both^{R3}	-	✓	sum	48.0M	**21.22**	**30.12**	26.53	22.06	**23.37**	26.59	19.82	12.54	**22.78**	12.6
19	par	both^{T3}	-	✓	sum	48.0M	20.35	28.61	25.94	21.22	22.12	25.19	19.36	11.99	21.85	13.6

* no normalization for dual-attention input, † `sum` merging has $\lambda = 0.3$ fixed,
R3 ASR is 3 steps ahead of ST, T3 ST is 3 steps ahead of ASR.

Detailed results

Wait-k: *ST-waits-for-ASR improves performance*, *ASR-waits-for-ST worsen*.

No	type	side	self	src	merge	params	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	independent (shared)					31.3M	19.40	27.77	24.65	19.93	21.53	24.24	18.19	10.99	20.84	14.2
2	independent					44.8M	20.11	28.18	25.61	20.76	21.83	25.45	18.45	11.31	21.46	12.6
3	independent++					51.2M	20.25	29.48	26.10	21.05	22.34	**26.71**	19.67	12.10	22.21	12.9
4	crx	st	-	✓	sum	46.4M	20.01	28.57	25.86	20.66	22.26	25.36	19.06	12.00	21.72	12.7
5	crx	both	-	✓	concat	51.2M	20.36	28.51	25.80	21.18	22.10	25.24	19.55	11.89	21.83	12.3
6	crx	both	-	✓	sum	48.0M	19.99	28.87	26.09	20.94	21.67	25.42	18.85	11.83	21.71	12.2
7	crx	both	✓	✓	concat	54.3M	20.07	28.73	26.01	20.93	22.59	25.60	19.08	12.46	21.93	12.4
8	crx	both	✓	✓	sum	51.2M	20.38	28.90	26.64	21.07	22.61	26.23	19.44	12.12	22.17	**12.1**
9	crx*	both	✓	-	sum	48.0M	19.72	27.96	25.49	20.52	21.56	25.01	18.53	11.33	21.26	12.8
10	crx*	both	✓	-	sum†	48.0M	18.62	27.11	24.41	19.73	20.47	24.49	17.23	11.09	20.39	12.8
11	crx*	both	✓	✓	sum	51.2M	19.54	28.17	25.68	20.95	21.55	24.77	18.76	11.28	21.34	12.3
12	par	st	✓	✓	concat	49.6M	20.57	28.84	26.08	20.85	22.11	25.70	19.36	11.90	21.93	13.0
13	par	both	-	✓	concat	51.2M	20.84	29.51	26.44	21.53	22.68	25.94	19.04	12.60	22.32	12.5
14	par	both	-	✓	sum	48.0M	20.85	29.18	26.38	**22.14**	22.87	26.49	19.70	12.74	22.54	12.7
15	par	both	✓	-	sum	48.0M	20.56	29.21	26.54	21.07	22.51	25.75	19.64	**12.80**	22.26	12.8
16	par	both	✓	✓	concat	54.3M	**21.22**	29.50	**26.66**	21.74	22.76	26.66	**20.25**	12.79	22.70	12.7
17	par	both	✓	✓	sum	51.2M	20.95	28.67	26.45	21.31	22.29	25.87	19.53	12.24	22.16	12.8
18	par	both^{R3}	-	✓	sum	48.0M	**21.22**	**30.12**	26.53	22.06	**23.37**	26.59	19.82	12.54	**22.78**	12.6
19	par	both^{T3}	-	✓	sum	48.0M	20.35	28.61	25.94	21.22	22.12	25.19	19.36	11.99	21.85	13.6

* no normalization for dual-attention input, † `sum` merging has $\lambda = 0.3$ fixed,
R3 ASR is 3 steps ahead of ST, T3 ST is 3 steps ahead of ASR.

Comparison to the state of the art

No	type	side	self	src	merge	epochs	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	Bilingual (Inaguma et al., 2020)					50	22.91	27.96	32.69	23.75	27.43	28.01	21.90	**15.75**	25.05	12.0
2	One-to-many (Gangi et al., 2019)						17.70	20.90	26.50	18.00	20.00	22.60	-	-	-	-
3	One-to-many (Gangi et al., 2019)						16.50	18.90	24.50	16.20	17.80	20.80	15.90	9.80	17.55	-
4	independent++					25	22.82	27.20	32.11	23.34	26.67	28.98	21.37	14.34	24.60	11.6
5	par	both	✓	✓	concat	25	22.74	27.59	32.86	23.50	26.97	29.51	21.94	14.88	25.00	11.6
6	par^{R3}	both	-	✓	sum	25	22.84	27.92	32.12	23.61	27.29	29.48	21.16	14.50	24.87	11.6
7	par++	both	-	✓	sum	25	**23.63**	**28.12**	**33.45**	**24.18**	**27.55**	**29.95**	**22.87**	15.21	**25.62**	**11.4**

Comparison to the state of the art

No	type	side	self	src	merge	epochs	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	Bilingual (Inaguma et al., 2020)					50	22.91	27.96	32.69	23.75	27.43	28.01	21.90	**15.75**	25.05	12.0
2	One-to-many (Gangi et al., 2019)						17.70	20.90	26.50	18.00	20.00	22.60	-	-	-	-
3	One-to-many (Gangi et al., 2019)						16.50	18.90	24.50	16.20	17.80	20.80	15.90	9.80	17.55	-
4	independent++					25	22.82	27.20	32.11	23.34	26.67	28.98	21.37	14.34	24.60	11.6
5	par	both	✓	✓	concat	25	22.74	27.59	32.86	23.50	26.97	29.51	21.94	14.88	25.00	11.6
6	par^{R3}	both	-	✓	sum	25	22.84	27.92	32.12	23.61	27.29	29.48	21.16	14.50	24.87	11.6
7	par++	both	-	✓	sum	25	**23.63**	**28.12**	**33.45**	**24.18**	**27.55**	**29.95**	**22.87**	15.21	**25.62**	**11.4**

- *Performance competitive to bilingual*, despite *simpler training recipe* and *fewer epochs*.

Comparison to the state of the art

No	type	side	self	src	merge	epochs	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	Bilingual (Inaguma et al., 2020)					50	22.91	27.96	32.69	23.75	27.43	28.01	21.90	**15.75**	25.05	12.0
2	One-to-many (Gangi et al., 2019)						17.70	20.90	26.50	18.00	20.00	22.60	-	-	-	-
3	One-to-many (Gangi et al., 2019)						16.50	18.90	24.50	16.20	17.80	20.80	15.90	9.80	17.55	-
4	independent++					25	22.82	27.20	32.11	23.34	26.67	28.98	21.37	14.34	24.60	11.6
5	par	both	✓	✓	concat	25	22.74	27.59	32.86	23.50	26.97	29.51	21.94	14.88	25.00	11.6
6	par^{R3}	both	-	✓	sum	25	22.84	27.92	32.12	23.61	27.29	29.48	21.16	14.50	24.87	11.6
7	par++	both	-	✓	sum	25	**23.63**	**28.12**	**33.45**	**24.18**	**27.55**	**29.95**	**22.87**	15.21	**25.62**	**11.4**

- *Performance competitive to bilingual*, despite *simpler training recipe* and *fewer epochs*.
- Largely *surpassing previous multilingual models* on MuST-C.

Comparison to the state of the art

No	type	side	self	src	merge	epochs	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	Bilingual (Inaguma et al., 2020)					50	22.91	27.96	32.69	23.75	27.43	28.01	21.90	**15.75**	25.05	12.0
2	One-to-many (Gangi et al., 2019)						17.70	20.90	26.50	18.00	20.00	22.60	-	-	-	-
3	One-to-many (Gangi et al., 2019)						16.50	18.90	24.50	16.20	17.80	20.80	15.90	9.80	17.55	-
4	independent++					25	22.82	27.20	32.11	23.34	26.67	28.98	21.37	14.34	24.60	11.6
5	par	both	✓	✓	concat	25	22.74	27.59	32.86	23.50	26.97	29.51	21.94	14.88	25.00	11.6
6	par^{R3}	both	-	✓	sum	25	22.84	27.92	32.12	23.61	27.29	29.48	21.16	14.50	24.87	11.6
7	par++	both	-	✓	sum	25	**23.63**	**28.12**	**33.45**	**24.18**	**27.55**	**29.95**	**22.87**	15.21	**25.62**	**11.4**

- *Performance competitive to bilingual*, despite *simpler training recipe* and *fewer epochs*.
- Largely *surpassing previous multilingual models* on MuST-C.
- Largest improvement on Portuguese (least data).

Thank you for your attention!

Code and pre-trained models are available:
`https://github.com/formiel/speech-translation`.