Speech Translation at Facebook

Juan Pino

FACEBOOK AI

Speech Translation

“Hello friends”

Speech-to-text

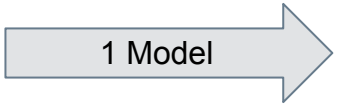
“Hola amics”

Speech-to-speech

What Speech Translation System?

End-to-End

“Hello friends”



“Hola amics”

- Multi-task

  (Tang et al. 2021)

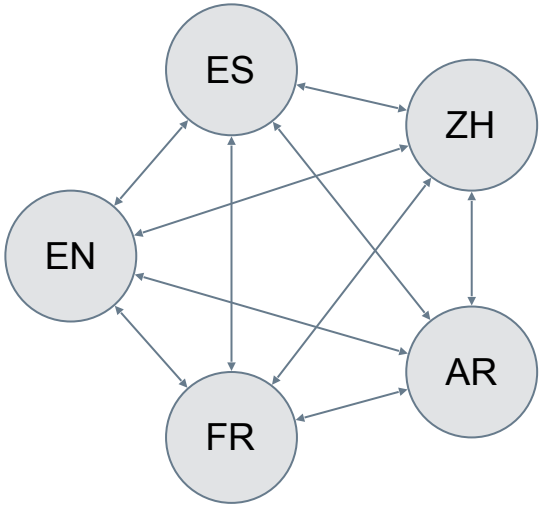
- Synchronous ASR/ST

  (Le et al. 2020)

Why End-to-End?

	Cascade	End-to-End
Model Size	✗	✓
Inference Latency	✗	✓
Compounding Errors	✗	✓
Preserve Input Audio Properties	✗	✓
Data	✓	✗

Multilingual



- Maintainability + GPUs
- Improve quality of some language pairs
- Audio: better transfer?

Local

- Privacy
- Inclusive of poor connectivity regions

Real-Time

- Generate translations with partial input

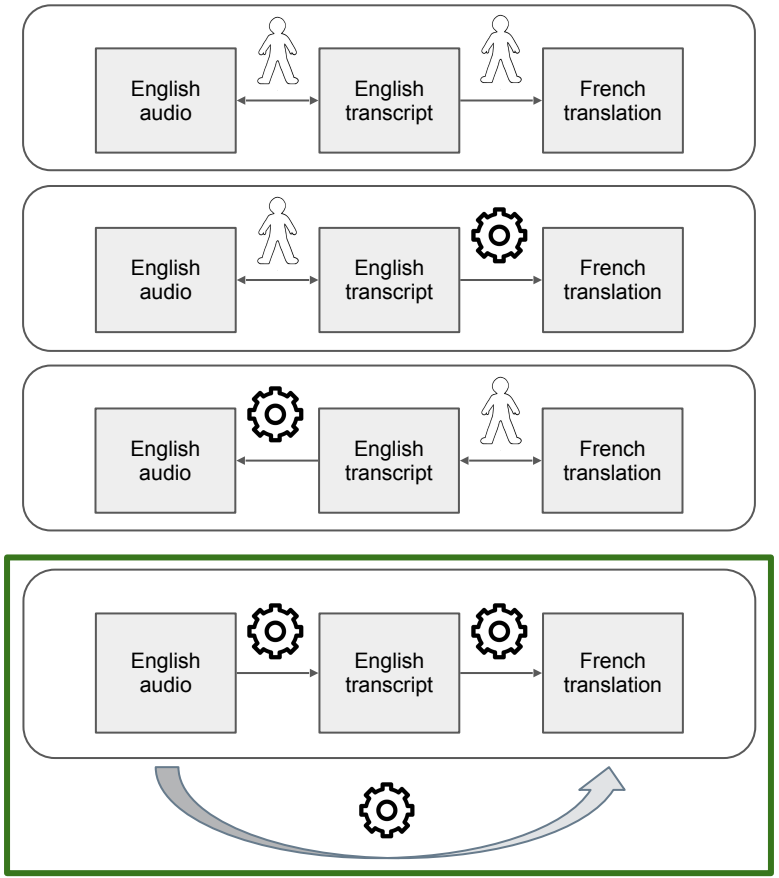
- Example application: simultaneous interpretation

Outline

- e2e: data scarcity challenge
  - [data augmentation] Self-Training for end-to-end ST
  - [pre-training] Multilingual ST with Efficient Finetuning of Pretrained Models (+ intro to CoVoST)
- simultaneous ST
  - SimulMT to SimulST
  - Streaming Simultaneous ST

Self-Training

Approach


English audio
English transcript
French translation
English audio
English transcript
French translation
English audio
English transcript
French translation
English audio
English transcript
French translation

Data Setup

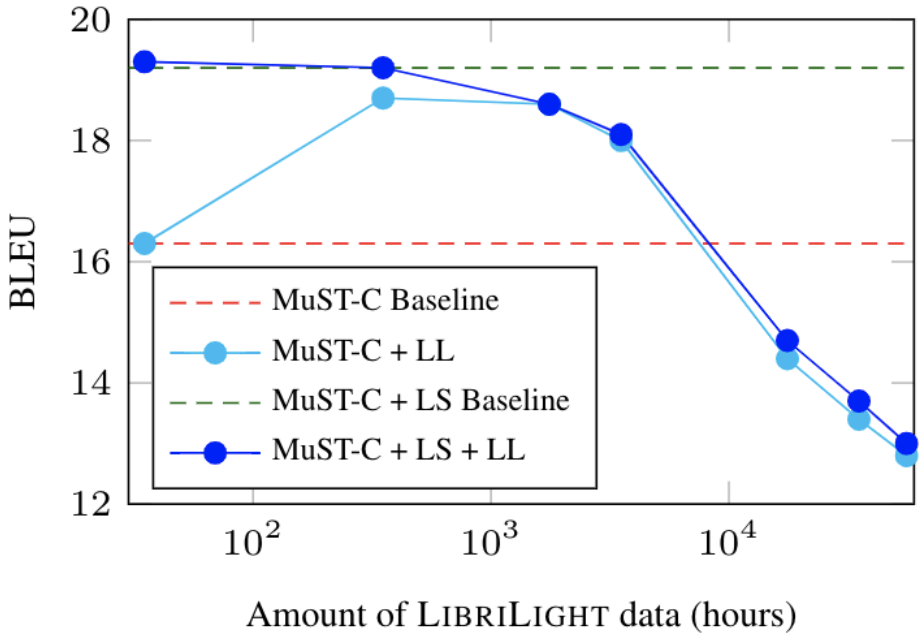
Open	Supervised	MuST-C	479/395 h
	Weakly Supervised	LibriSpeech (LS)	960 h
	Unsupervised	LibriLight (LL)	56k h
Internal	Supervised + Weakly Supervised	FB Video + Open	30k h
	Unsupervised	FB Video	255k h

Models

Model	# Parameters
LSTM	13.5M
VggT	260M
VggTLarge	435M

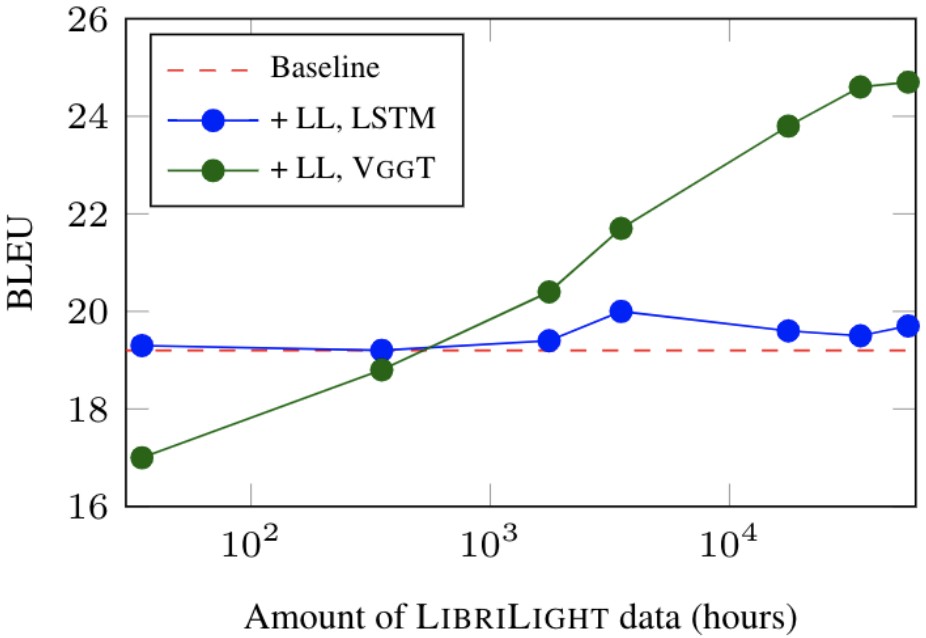
- Scale model size with data
- Transformer for training speed

Initial Results



- Improvement over weak baseline
- No improvement over strong baseline
- Degradation with amount of data

Improving Stronger Baseline (fine-tuning, larger architecture)



- LSTM: small improvement over strong baseline
- VggT: large improvement over strong baseline

Scaling up Data/Model Size

Data	VggT (260M)	VggTLarge (435M)
MuST-C + LS + 17k h LL	23.8	23.7
MuST-C + LS + 35k h LL	24.6	25.6

Main Results

Domain	Language	Data	BLEU
MuST-C	En-Fr	SoTA Baseline*	34.05
		MuST-C + LibriSpeech	26.2
		+35k hours LibriLight	**34.5**
	En-De	SoTA Baseline	22.11
		MuST-C + LibriSpeech	19.5
		+35k hours LibriLight	**25.2**
FB Videos	En-Fr	Baseline	20.3
		+96k hours videos	**21.6**
	En-Es	Baseline	18.5
		+96k hours videos	**19.9**

***** (Indurthi et al., ICASSP 2020)

Ablations: Encoder Pre-Training (via ASR)

Data	ASR Pre-training	Pseudo-Labeling
MuST-C + LS	19.8 (25.2)	3.7
+3523 h LL	20.8 (22.7)	21.7
+17,607 h LL	21.0 (40.3)	23.9

- Both methods are effective.
- Pseudo-labeling provides more gains.
- WER degrades with more data but still helps translation task.

Self-Training

Data	Pseudo-Labeling Model	Model	BLEU
MuST-C	N/A	LSTM	16.3
+3523 h LL	Cascade	LSTM	**20.8**
	LSTM		18.5
	VggT		20.6
MuST-C + LS	N/A	LSTM	19.2
+17607 h LL	Cascade	VggT	23.8
	LSTM		20.7
	VggT		**24.5**

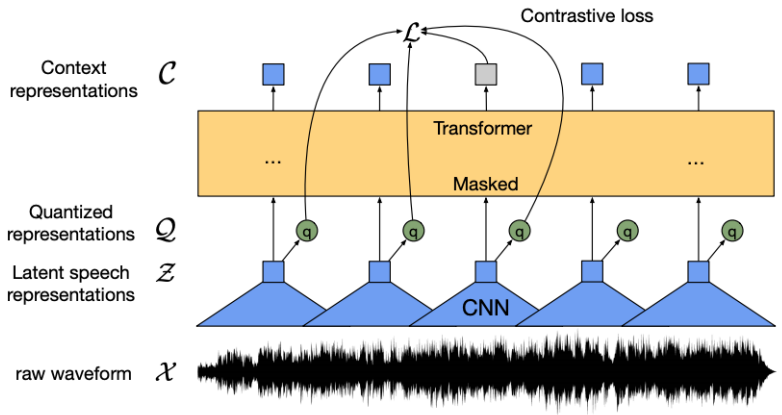
- All methods work, including pure self-training.
- The VggT pseudo-labeling model is competitive with the cascade.

Conclusions

- Fine-tuning and scaling up model and data size are essential.
- Decoder pre-training is important.
- Pseudo-labeling via end-to-end model is possible after bootstrap phase.

Pre-Training

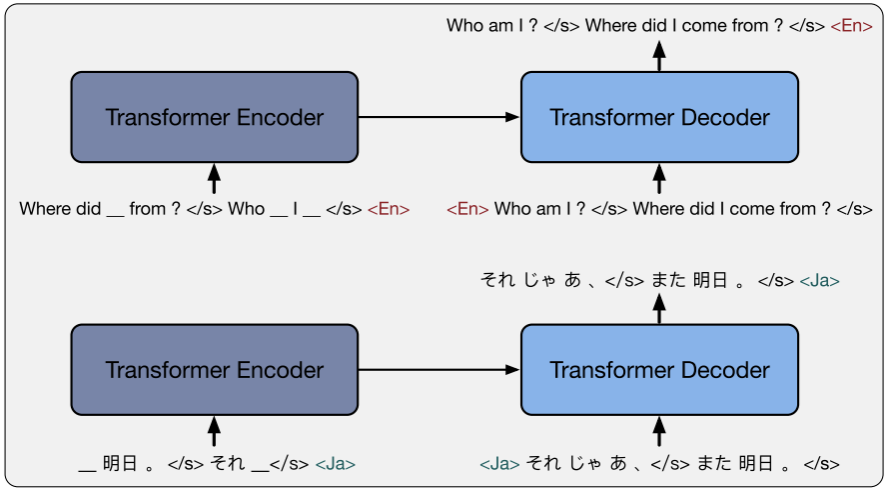
Wav2vec 2.0



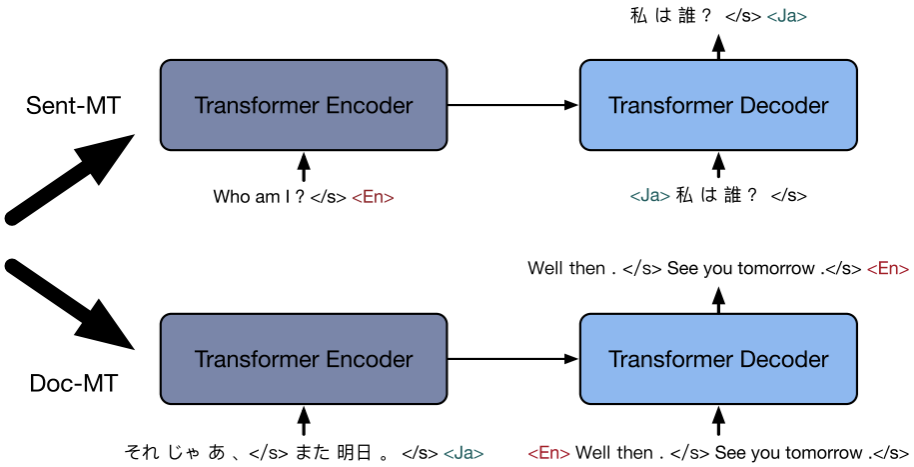
- Context representations learned directly on raw input
- SoTA results on low and very low resource LibriSpeech
- Competitive results on full LibriSpeech

(Baevski et al., 2020)

mBART



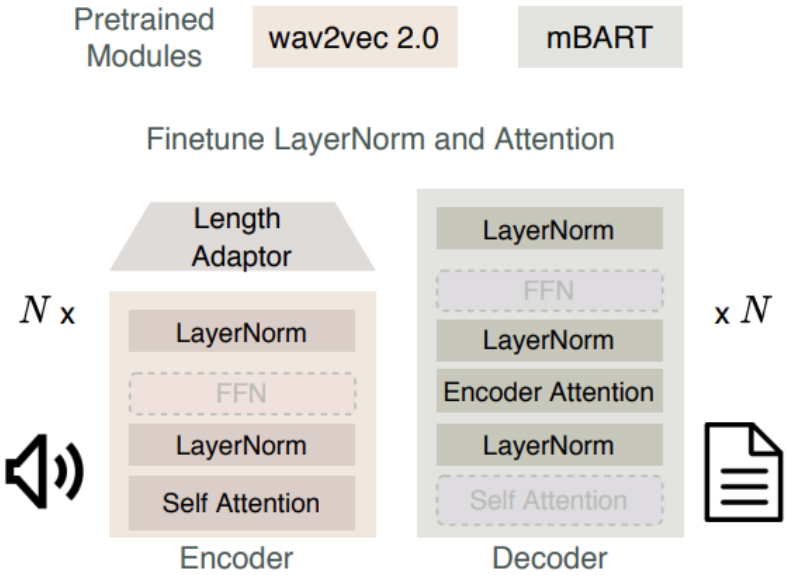
Multilingual Denoising Pre-Training (mBART)



Fine-tuning on Machine Translation

(Liu et al., 2020)

Approach



- pre-training
  - encoder: wav2vec 2.0
  - decoder: mBART
  - note: mBART **fine-tuned on MT data**
- adapter (similar to traditional model)
- only fine-tune LNA

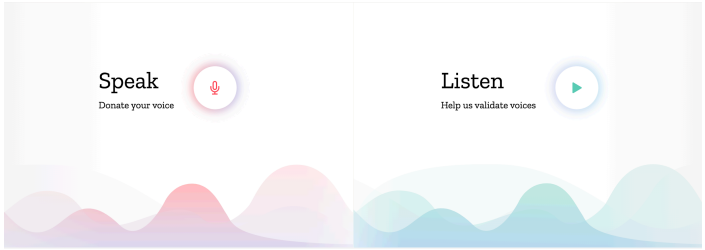
Data

- Europarl-ST (non-English, zero-shot)

- CoVoST 2 (multilingual, low resource pairs)

CoVoST

- Built on top of Common Voice.
- 22 languages (21-English, English-15)
- 2800 hours, 78k speakers
- https://github.com/facebookresearch/covost



CoVoST (Next Versions)

- Latest CV release: 60 languages/9000+ hours
- Plan for support for 49 languages, including low resource languages such as Kinyarwanda and dialects such as Kabyle

- Inject CoVoST translations into Common Voice crowdsourcing platform: s2s alignments

Main Results (many2one)

Config	Avg Test BLEU
Multilingual Baseline	7
LNA (170M)	12.5

Main Results (one2many)

Config	Avg Test BLEU
Multilingual Baseline	18.1
LNA (170M)	22.5

Zero-Shot (Source)

Config	Pt-En Test BLEU
ASRPT + Multi	4.4
Supervised	6.3
LNA (170M)	8.2

- Fine-tune on Fr/De/Es/Ca/It-En
- Test on Pt-En

Zero-Shot (Target)

Config	Pt-En Test BLEU
ASRPT + Multi	0
Supervised	31.9
LNA (170M)	29.2

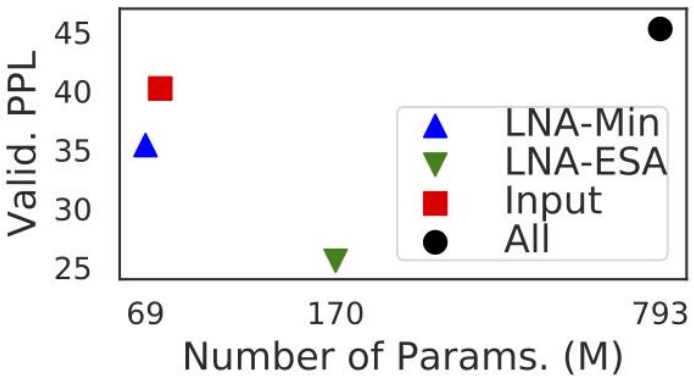
- Fine-tune on En-De/Fa/Tr/Zh
- Test on En-Ja

Zero-Shot (source/target)

Source	Target: De	En	Es	Fr	It	Pt
De		12.8/**20.6**	10.2/**13.8**	11.6/**14.9**	6.6/**8.6**	10.4/**13.0**
En	13.1/**22.5***		23.1/**32.3***	22.1/**30.0***	14.9/**21.5**	20.7/**28.4**
Es	9.2/**12.1**	18.9/**26.0**		19.0/**21.8**	13.3/**15.4**	20.0/**21.9**
Fr	9.8/**13.6**	19.8/**27.9***	18.6/**21.7**		13.8/**15.2**	19.7/**21.4**
It	10.1/**11.9**	19.8/**25.6**	18.8/**20.8**	19.1/**20.0***		19.8/19.2
Pt	9.0/**11.4**	19.0/**24.1**	19.8/19.6	18.1/**18.6**	15.6/**16.1**	

- Fine-tune on X-En and En-Y Europarl-ST
- Test on X-Y

Efficient Fine-Tuning



- Data: De-En CoVoST
- Contrast All, Input (feature encoder), LNA (Min: no self-attention)

Conclusions

- wav2vec 2.0 + mBART pretraining is effective
- LNA fine-tuning is efficient
- Approach also effective for zero-shot

Real Time

Offline vs Real-Time

- Offline: read sentence/audio segment and translate

- Real-Time:
  - Re-Translation: re-translate entire input whenever new input is read
  - Simultaneous: translate with partial input, typically no backtracking

Offline

Heute spreche ich zu Ihnen über Energie und Klima.

I'm going to talk today about energy and climate.

Re-Translation

Heute

Today

Simultaneous

Heute

Evaluation

- Quality: BLEU
- Latency: Average Lagging (AL)

Average Lagging

$$\mathrm{AL}_g(\mathbf{x}, \mathbf{y}) = \frac{1}{\tau_g(|\mathbf{x}|)} \sum_{t=1}^{\tau_g(|\mathbf{x}|)} g(t) - \frac{t-1}{r}$$

Target token position when all input is read

Number of source words read when generating token in position t (“lagging”)

Ideal system policy

Average Lagging (Speech Input)

$$\mathrm{AL} = \frac{1}{\tau(|\boldsymbol{X}|)} \sum_{t=1}^{\tau(|\boldsymbol{X}|)} d(t) - \frac{|\boldsymbol{X}|}{|\boldsymbol{Y}^*|} \cdot T_s \cdot (t-1)$$

Target token position when all **audio** input is read

Duration of audio input read when generating token in position t

~~Duration of audio~~ **Time to** generate token in position t

Ideal system policy

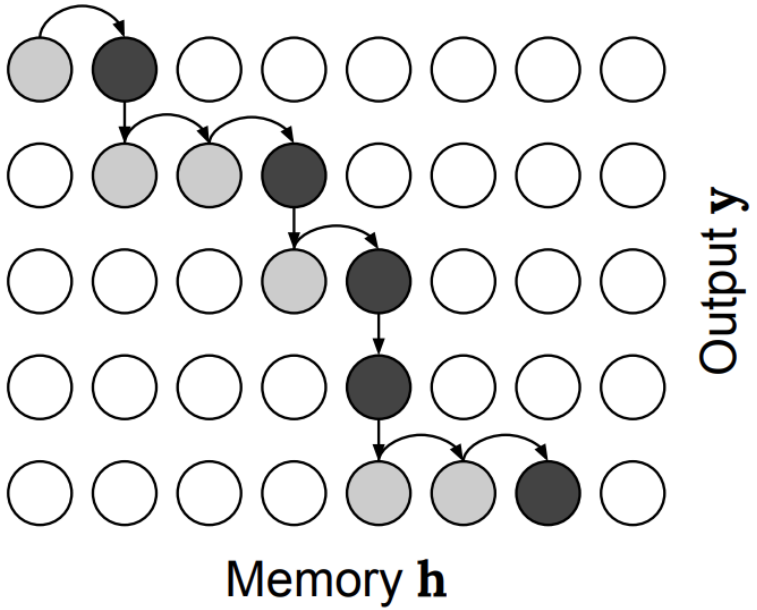
Simultaneous Translation Policies

- Wait-k (rule-based)
  - read k tokens, write 1 token, read 1 token, etc.

- Monotonic attention (model-based)
  - read-write probability determined by attention a(s, h)

Monotonic Attention

```
while y_{i-1} is not EOS
  for j = t_{i-1} to T
    e_{i,j} = a(s_{i-1}, h_j)
    p_{i,j} = sigmoid(e_{i,j})
    if WRITE
      c_i = h_j
      t_i = j
      break
  generate y_i
  i = i + 1
```

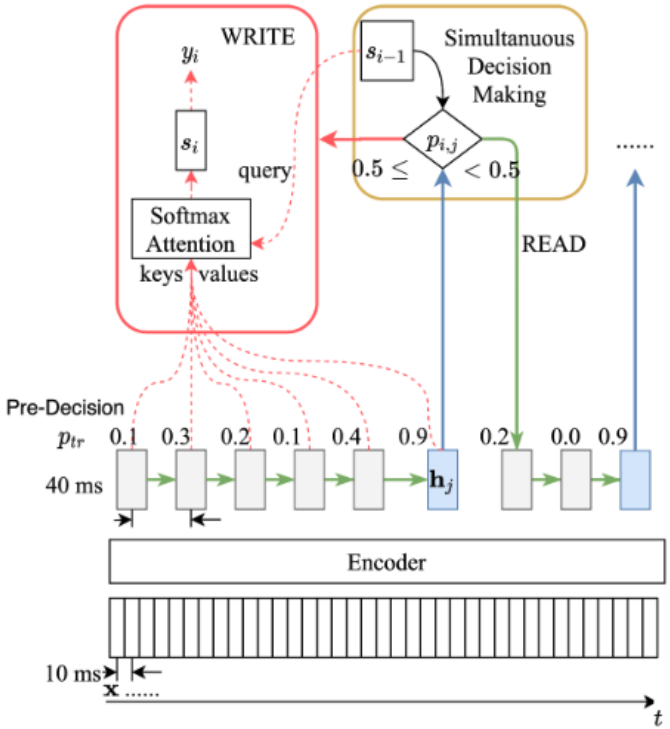


(Raffel et al., 2017)

SimuI ST: How to Group Frames ?

Pre-Decision

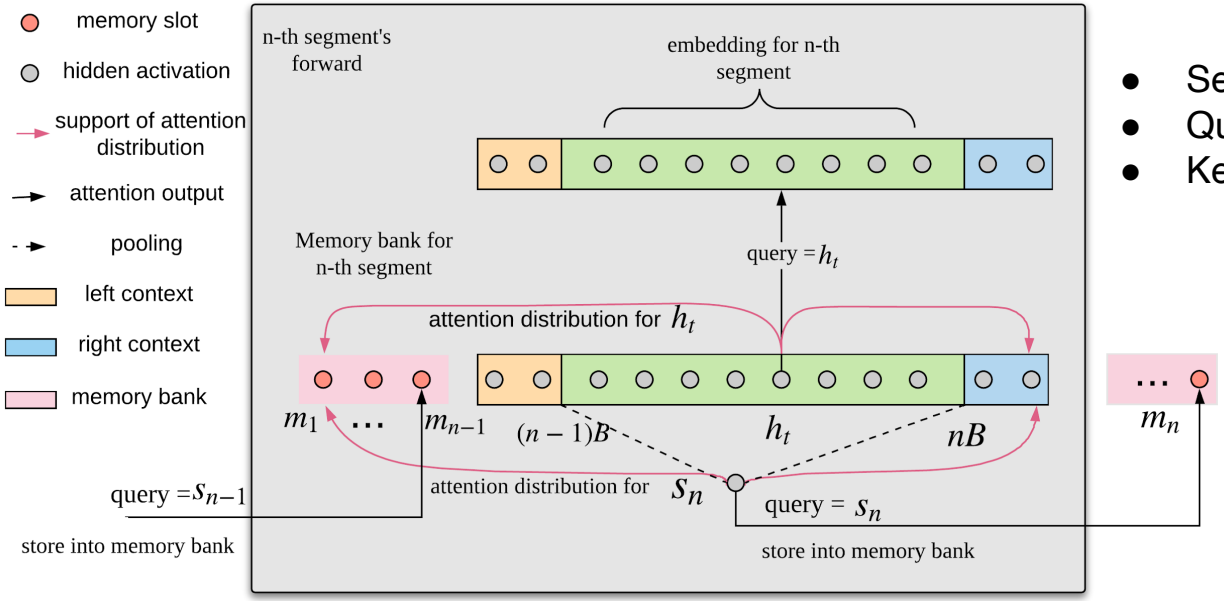
- Fixed: 280ms ~ word
- Flexible: oracle word (not phoneme) boundaries



Streaming vs Simultaneous

- Previous approach is simultaneous
- Not streaming: transformer encoder
- Solution: memory-augmented transformer encoder

Augmented Memory Transformer Encoder



- Segment with left and right context
- Query: L, C, R, s
- Key, Value: M, L, C, R

Ongoing/Future Work

- Larger-scale experiments with streaming models
- Expanding CoVoST + Mining
- Expanding pre-training work to more languages
- S2S

Thank You

- Many programs: research/engineering internships, fellowships, FT, etc.
- IWSLT shared task: https://iwslt.org/2021/simultaneous
- https://github.com/pytorch/fairseq/tree/master/examples/speech_to_text

FACEBOOK AI