How to get the model to do what we want:

Simultaneous Speech Translation

Dominik Macháček
11/4/2024

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Speech Translation

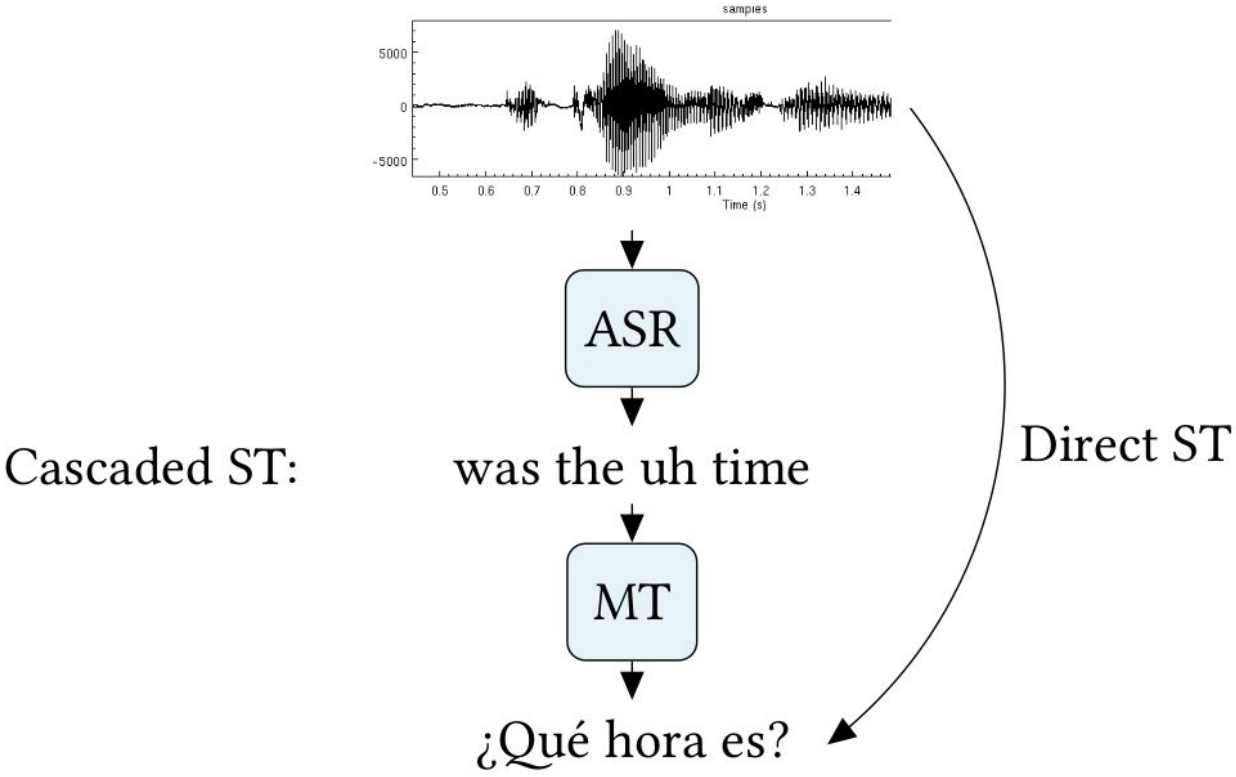
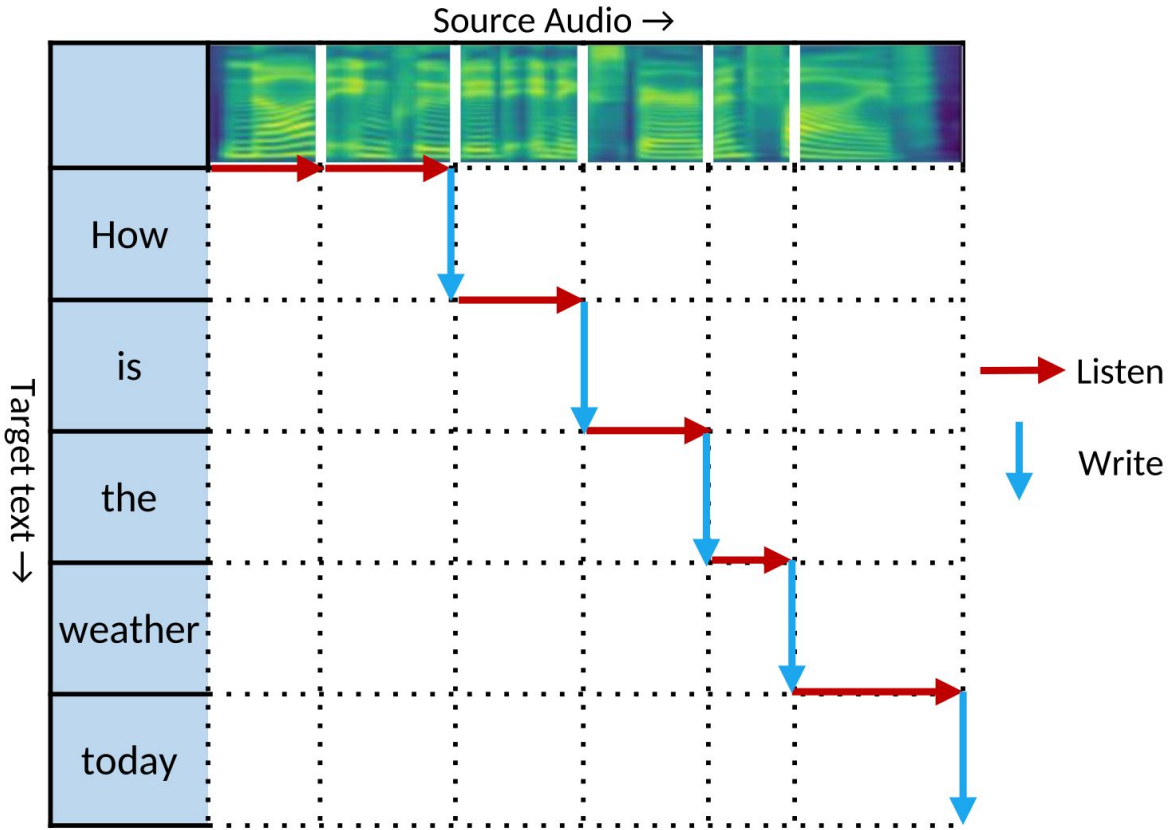


Image from the ELITR project, by Barry Haddow

Simultaneous Speech Translation



Simultaneous speech translation

- **Simultaneous** = Live = Real-Time = Low-latency = Incremental
  - Source available continuously, one **chunk** at a time
  - The **chunk** can be:
    - audio segment … in the direct speech-to-text translation or transcription = ASR
    - or word (text) produced by incremental ASR … in a cascaded system = ASR + **MT**
  - Provide the target "at the same time" as the source is being produced

    = simultaneously = with a small additive delay

In the European Parliament:

-> English original source

-> English-to-Czech Sim. Interpreting

-> English-to-German Sim. Intp.

Simultaneous speech translation

- **Simultaneous** = Live = Real-Time = Low-latency = Incremental
  - Source available continuously, one **chunk** at a time
  - The **chunk** can be:
    - audio segment … in the direct speech-to-text translation or transcription = ASR
    - or word (text) produced by incremental ASR … in a cascaded system = ASR + **MT**
  - Provide the target "at the same time" as the source is being produced

    = simultaneously = with a small additive **delay**

In the European Parliament:

-> English original source

-> English-to-Czech Sim. Interpreting

-> English-to-German Sim. Intp.

Challenges:

... tell me

Challenge: Waiting for the context

Chinese source:

jīngfāng	xiàzhōu	jiāng	duì	bù fèn	shè àn	rén yuán	tí qǐ gōng sù
警方	下周	将	对	部分	涉案	人员	提起 公诉
police	next week	will	for	part	involved	people	**accuse**

Simultaneous intp.:

Next week, police

will

accuse some of the people involved in the case.

Source: https://aclanthology.org/2020.emnlp-main.178.pdf

Challenges: Most of parallel data are for translation

Chinese source:

jǐngfāng	xiàzhōu	jiāng	duì	bù fèn	shè àn	rén yuán	tí qǐ gōng sù
警方	下周	将	对	部分	涉案	人员	提起 公诉
police	next week	will	for	part	involved	people	**accuse**

English "offline" translation: … no problem with re-ordering

Police will accuse some of the people involved in the case next week.

But in simultaneous: … face the word order diff. + wait / or guess and risk being wrong

Next week, police will ...[long waiting]… accuse some of the people involved in the case.

Source: https://aclanthology.org/2020.emnlp-main.178.pdf

Challenges:

- Word orders: wait or translate / re-translate with every new chunk
- Quality
- Latency
- Stability
- Model
  - + Training
  - + Data
  - + Decoding
- Practical

Simultaneous approaches

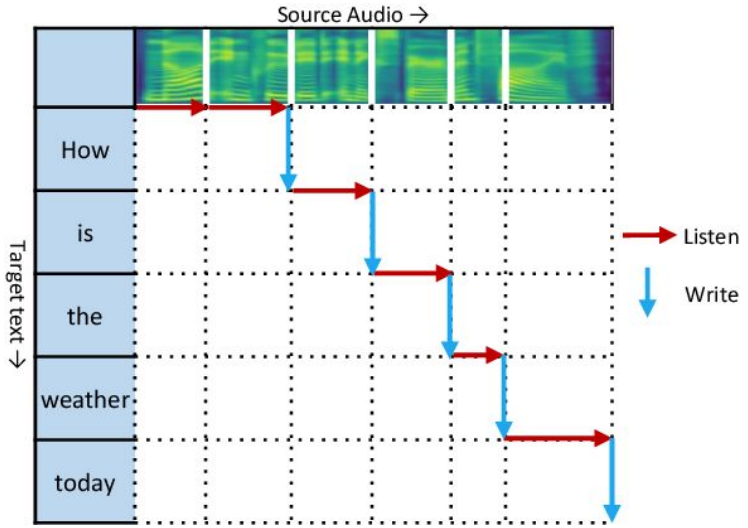
Re-Translation vs. Wait and append

- Re-translate from beginning of sentence each time: **rewrite + append**
- Latency vs stability. **Top quality.**

Source	Output								Erasure
1: Neue	New								-
2: Arzneimittel	New	Medicines							0
3: könnten	New	Medicines							0
4: Lungen-	New	drugs	may	be	lung				1
5: und	New	drugs	could	be	lung	and			3
6: Eierstockkrebs	New	drugs	may	be	lung	and	ovarian	cancer	4
7: verlangsamen	New	drugs	may	slow	lung	and	ovarian	cancer	5
Content Delay	1	4	6	7	7	7	7	7	

Source: [Arivazhagan et al., 2020]

- Alternates between reading from ASR and translating: **no rewrites, only append**
- Latency vs. quality. **Top stability.**



Source: [Ren et al., 2020]

Stability in Re-Translation

How to make re-translation more stable?

Baseline (“standard” offline MT)

Source	Output								Erasure
1: Neue	New								-
2: Arzneimittel	New	Medicines							0
3: könnten	New	Medicines							0
4: Lungen-	New	drugs	may	be	lung				1
5: und	New	drugs	could	be	lung	and			3
6: Eierstockkrebs	New	drugs	may	be	lung	and	ovarian	cancer	4
7: verlangsamen	New	drugs	may	slow	lung	and	ovarian	cancer	5

Stability measure: **13** erasures for **8** generated tokens = **1.625**

Improvement

Source		Erasure
1: Neue	New	-
2: Arzneimittel	New drugs	0
3: könnten	New drugs may	0
4: Lungen-	New drugs may lung	0
5: und	New drugs may lung and	0
6: Eierstockkrebs	New drugs may lung and ovarian cancer	0
7: verlangsamen	New drugs may slow lung and ovarian cancer	4

4 erasures for **8** generated tokens = **0.5**

Example: [Arivazhagan et al., 2020]. The method first proposed by …TBD.

How to make re-translation more stable?

Baseline (“standard” offline MT)

Source	Output	Erasure
1: Neue	New	-
2: Arzneimittel	New Medicines	0
3: könnten	New Medicines	0
4: Lungen-	New drugs may be lung	1
5: und	New drugs could be lung and	3
6: Eierstockkrebs	New drugs may be lung and ovarian cancer	4
7: verlangsamen	New drugs may slow lung and ovarian cancer	5

Stability measure: 13 erasures for **8** generated tokens = 1.625

Improvement

Source		Erasure
1: Neue	New	-
2: Arzneimittel	New drugs	0
3: könnten	New drugs may	0
4: Lungen-	New drugs may lung	0
5: und	New drugs may lung and	0
6: Eierstockkrebs	New drugs may lung and ovarian cancer	0
7: verlangsamen	New drugs may slow lung and ovarian cancer	5

4 erasures for **8** generated tokens = **0.5**

Learn this: Proportional prefix training

Full	Source	Die Führungskräfte der Republikaner rechtfertigen ihre Politik mit der Notwendigkeit , den Wahlbetrug zu bekämpfen [*15 tokens*]
	Target	Republican leaders justified their policy by the need to combat electoral fraud [*12 tokens*]
Prefix	Source	Die Führungskräfte der Republikaner rechtfertigen [*5 tokens*]
	Target	Republican leaders justified their [*4 tokens*]

Table 2: An example of proportional prefix training. Each example in the minibatch has a 50% chance to be truncated, in which case, we truncate its source and target to a randomly-selected fraction of their original lengths, 1/3 in this example. No effort is made to ensure that the two halves of the prefix pair are semantically equivalent.

1. Train a standard offline MT
2. Finetune on 1:1 mix of full sent. pairs and src-target prefixes
3. Create the prefixes from the length proportion,
4. do not care about the parallel words in the truncated suffix => anticipation
5. Measure the MT quality and erasures
6. Select a suitable trade-off

Example: [Arivazhagan et al., 2020]. The method first proposed by Niehues et al., 2018.

My results: learning curves

BLEU vs. steps

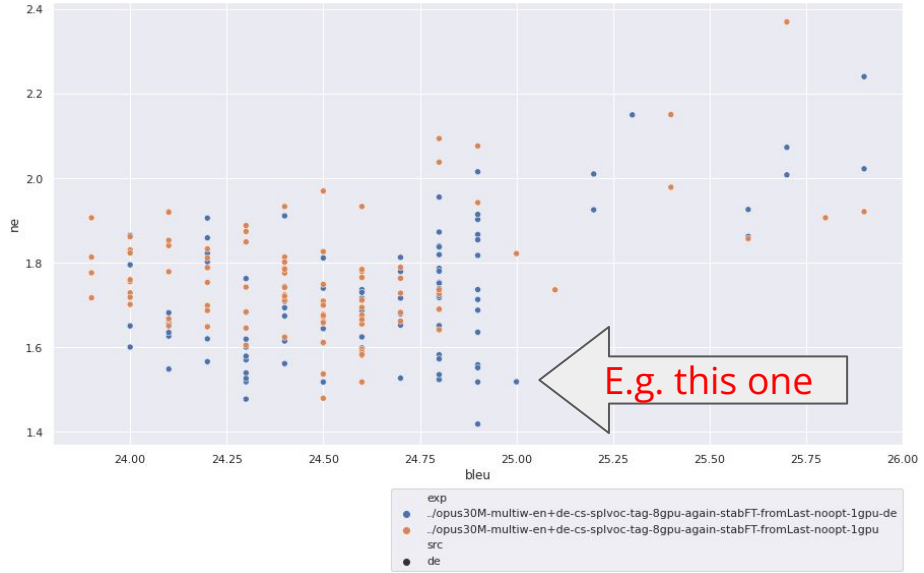
Colors are src-tgt variants, from the top:
En->Cs, En+De->Cs (multi-sourcing), De->Cs

Stability vs. steps

De->Cs, En->Cs

Select a checkpoint: quality-stability trade-off

Stability vs. MT quality (checkpoints for De-Cs)



	En		De	
checkpoint	BLEU	NE	BLEU	NE
starting	33.2	1.77	25.9	3.15
selected	33.0	1.21	25.0	1.52
diff	-0.2	-40%	-0.9	-52%

Table 6.13: The results of fine-tuning for stability, on ESIC dev. NE stands for "Normalized Erasure" (Arivazhagan et al., 2020b), measure of stability of re-translating simultaneous MT.

Results … for En->Cs, -0.2 BLEU, 40% higher stability
… for De->Cs, -0.9 BLEU, 52% higher stability

Other simultaneous problems and solutions

Other SST problems and solutions (briefly)

- How to learn when to wait and translate?

Other SST problems and solutions (briefly)

- How to learn when to wait and translate?
  => RL agent (outdated),
  => or simultaneous streaming policies (in the next lessons)

Other SST problems and solutions (briefly)

- How to train simultaneous encoder-decoder effectively? On all the prefixes at once?
  => Encoder with monotonic look-back attention

Monotonic Infinite Lookback Attention for Simultaneous Machine Translation

Naveen Arivazhagan* **Colin Cherry*** **Wolfgang Macherey** **Chung-Cheng Chiu**

Semih Yavuz **Ruoming Pang** **Wei Li** **Colin Raffel**

Google

Other SST problems and solutions (briefly)

- How to **continue translation** with the previous prefix?

Other SST problems and solutions (briefly)

- How to **continue translation** with the previous prefix?
  => autoregressive decoding can start with any tgt. prefix

Other SST problems and solutions (briefly)

- How to suggest the **target terminology**?

Other SST problems and solutions (briefly)

- How to suggest the **target terminology**?
  Prompts.
  [OpenAI Whisper documentation:]

```
# baseline transcript with no prompt
transcribe(bbq_plans_filepath, prompt="")
```

"Hello, my name is Preston Tuggle. I'm based in New York City. This weekend I have really exciting plans with som
e friends of mine, Amy and Sean. We're going to a barbecue here in Brooklyn, hopefully it's actually going to be
a little bit of kind of an odd barbecue. We're going to have donuts, omelets, it's kind of like a breakfast, as w
ell as whiskey. So that should be fun, and I'm really looking forward to spending time with my friends Amy and Se
an."

While Whisper's transcription was accurate, it had to guess at various spellings. For example, it assumed the friends' names were spelled Amy and Sean rather than Aimee and Shawn. Let's see if we can steer the spelling with a prompt.

```
# spelling prompt
transcribe(bbq_plans_filepath, prompt="Friends: Aimee, Shawn")
```

"Hello, my name is Preston Tuggle. I'm based in New York City. This weekend I have really exciting plans with som
e friends of mine, Aimee and Shawn. We're going to a barbecue here in Brooklyn. Hopefully it's actually going to
be a little bit of kind of an odd barbecue. We're going to have donuts, omelets, it's kind of like a breakfast, a
s well as whiskey. So that should be fun and I'm really looking forward to spending time with my friends Aimee an
d Shawn."

Success!

Other SST problems and solutions (briefly)

- How to learn when to wait and translate?
  => RL (outdated), or simultaneous streaming policies (some other time)
- How to train simultaneous encoder-decoder effectively?
  => monotonic look-back attention in the encoder
- How to continue translation with the previous prefix?
  => autoregressive decoding can start with any tgt. prefix
- MT gives **too long** targets, the users in real-time need shorter synonyms.
  => filter a parallel corpus for the shorter src-tgt pairs, train on them
- MT is verbose and literal, but too complicated to perceive
  => style transfer, learn e.g. on the simultaneous interpreting data
  ... or synthesize them
- How to suggest the target terminology? Whisper model with prompting.
- Some other time: speech-to-text tutorial, interactive demo,
  Discussion in 99 languages

Some other time

- Live interactive **demo** – ELITR, Whisper-Streaming
  Live speech src. in 99 languages, translation into 43 langs.
- Speech-to-text models (= like LLMs with speech input)
- Simultaneous streaming policies

Summary

- You learned what is the Simultaneous Speech Translation
- What are its challenges
- You learned how to stabilize re-translation:
  finetune on prefixes

See demo next time!