Multi-Source Simultaneous Speech Translation

Dominik Macháček
machacek@ufal.mff.cuni.cz
March 4 2024

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

Monday seminar: Multi-Source Sim. Speech Translation

We investigate the opportunity to use multiple parallel speech signals — the original and simultaneous interpreting — as sources for translation to achieve higher quality. We create an evaluation set ESIC (Europarl Simultaneous Interpreting Corpus). We analyze the challenges of simultaneous interpreting when used as an additional parallel source. Then, we investigate the robustness of multi-sourcing to transcription errors and assess the reliability of machine translation metrics when evaluating simultaneous speech translation. Last but not least, we demonstrate Whisper-Streaming, our tool that enables real-time processing of large offline speech-to-text models.

Outline

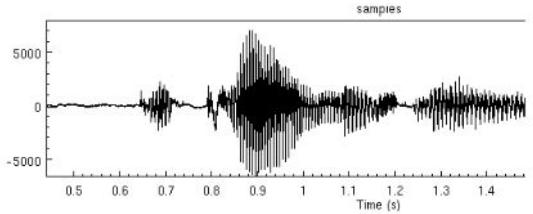
1) Long story short
2) Live Demo
3) Motivation
4) Specification
5) Data
6) Analysis of Interpreting
7) Multi-Sourcing Robustness
8) SST Human Evaluation in Simultaneous Mode
9) Whisper-Streaming
10) Summary
11) Next plans: InCroMin project

Outline ... references to dissertation

1)	Long story short	... Chapter 1
2)	Live Demo	... Section 8.4.5
3)	Motivation	... Chapter 2
4)	Specification	... Chapter 3
5)	Data	... Chapter 4
6)	Analysis of Interpreting	... Chapter 5
7)	Multi-Sourcing Robustness	... Chapter 6
8)	SST Human Evaluation in Simultaneous Mode	... Chapter 7, Section 8.2
9)	Whisper-Streaming	... Section 8.4
10)	Summary	... Chapter 9
11)	Next plans: InCroMin project	

1) Long Story Short

Speech Translation



ASR

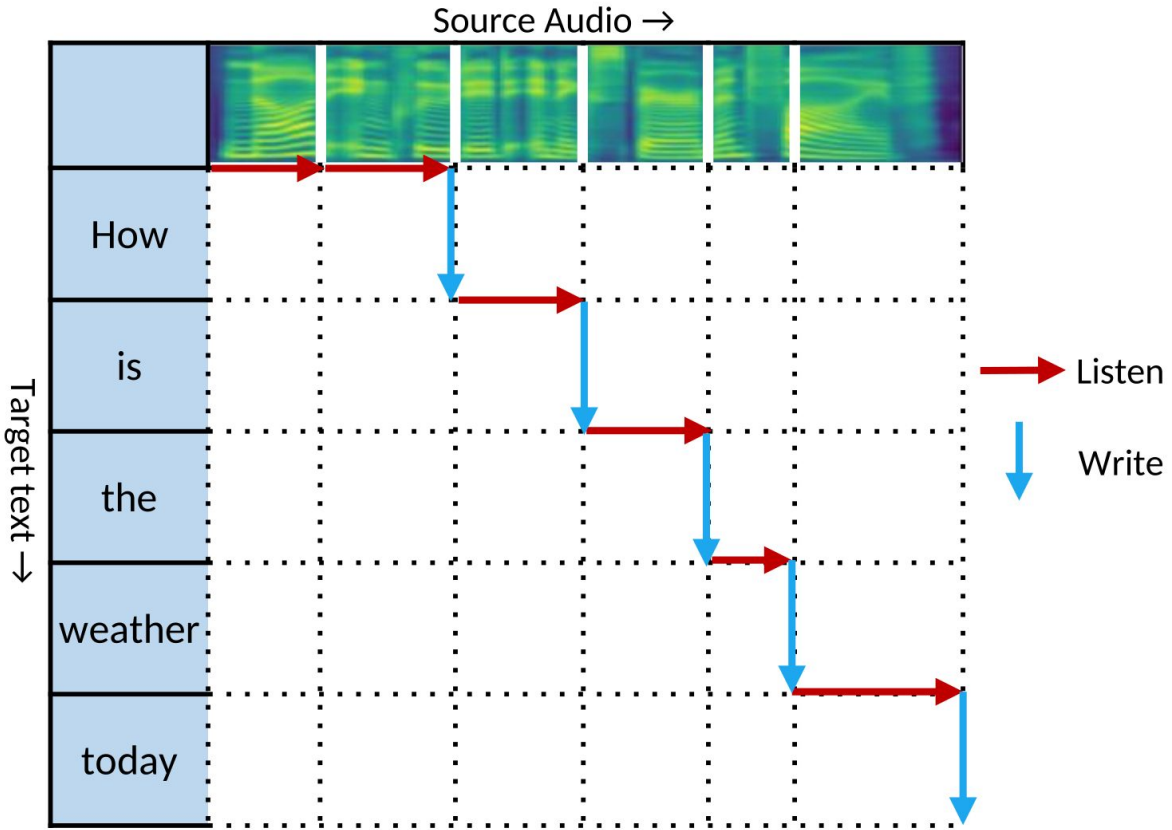
Cascaded ST: was the uh time

MT

¿Qué hora es?

Direct ST

Simultaneous Speech Translation

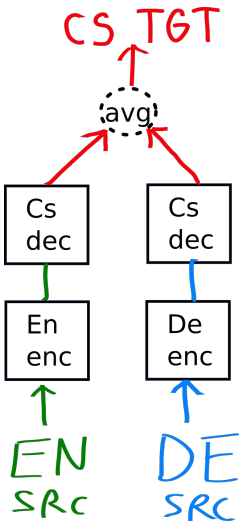


Multi-Source Simultaneous Speech Translation

En SRC: ...Mr Baş, the outgoing president of Eurosci

De SRC: ...Herrn Basch, den scheidenden Präsidenten der EUROSAI

Cs TGT: ...pana Başe, dosluhujícího prezidenta EUROSAI



...from the Source AND Interpreting

Challenges of Multi-Sourcing from Orig. + Interpreting:

- Quality
- Latency
- Alignment of sources
  - Time shift
  - Interpreting not literal
  - Sentence segmentation
- Balance costs and benefits
- Data
- Model
- Evaluation

Plan:

- Text-to-text MT for cascaded SST
- Supervised multi-sequence to sequence
- Synthetic training data
- Across sentence boundaries
- + Simultaneous mode

Results:

- Evaluation data
- Analysis of interpreting
- Multi-source model in simplified setup
- Background for the next steps:
  - Human + Metrics SST Evaluation
  - Whisper-Streaming

2) Live Speech-to-Text Demo

Demo on your device

- Click quest.ms.mff.cuni.cz/elitr/demo/
- ASR = transcript in orig. language, EN = translation into English, other columns: En->X MT
- **Whisper-Streaming:** voice activity controller + lang. detection + Whisper + streaming with LocalAgreement-2, wrapped in the ELITR pipeline

3) Motivation

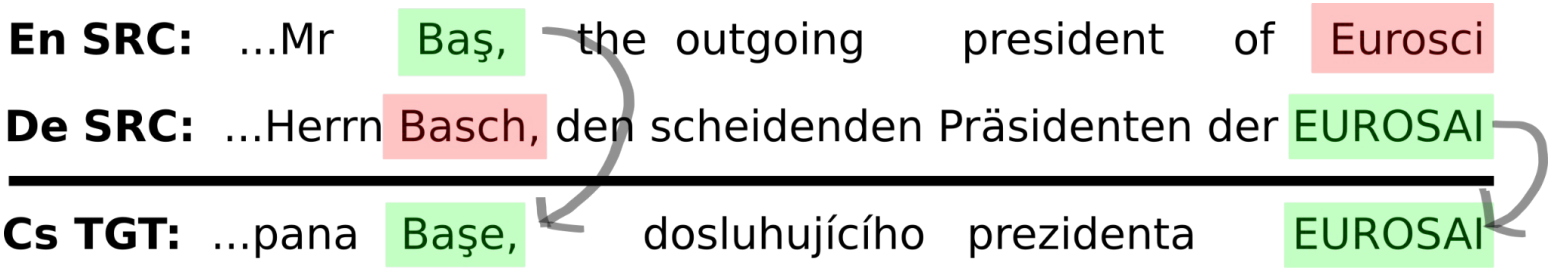
Benefits and Risks of Source+Interpreter

1) **Quality:**

Disambiguation

Schloss + castle/lock alien + Fremde/Außerirdischer [foreigner/extraterrestrial]

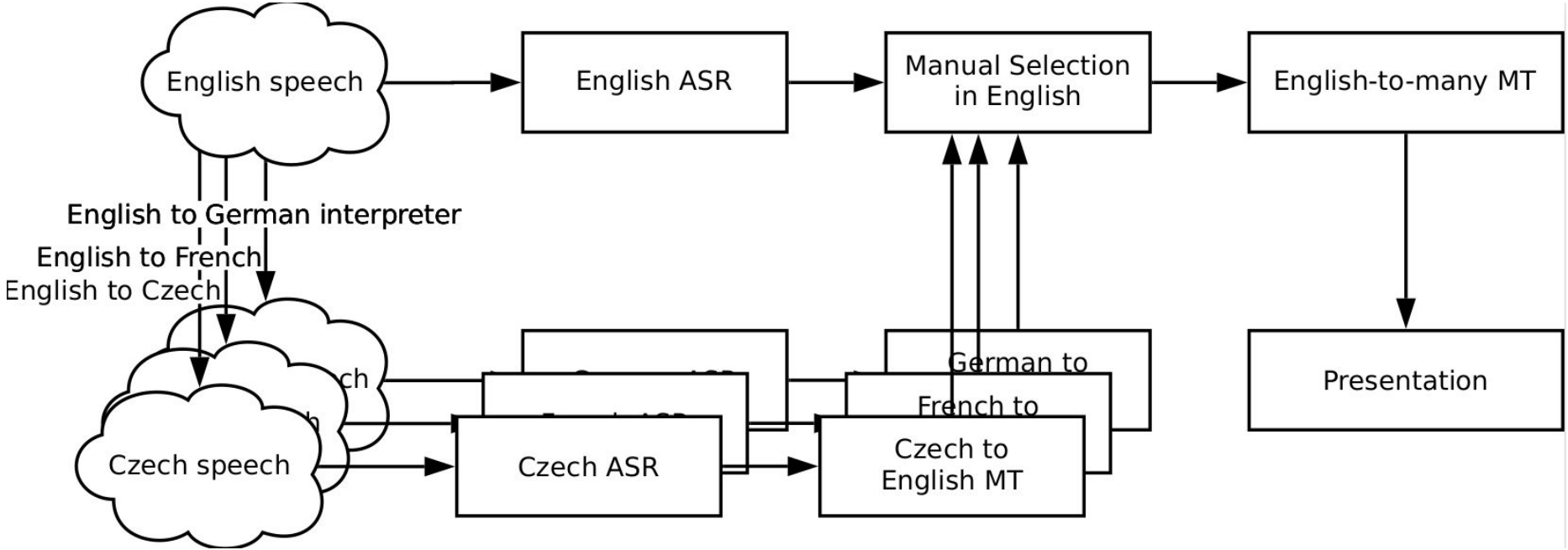
Complementary ASR errors:



Ref: Robustness of Multi-Source MT to Transcription Errors, ACL 23 Findings

Benefits and Risks of Source+Interpreter

2) No human interaction for detecting and switching the optimal source



[Bojar et al., 2021] Operating a Complex SLT System with Speakers and Human Interpreters

Benefits and Risks of Source+Interpreter

3) Possibly **best from both** options [Macháček et al., 2021, Lost in Interpreting...]
   - **Source:** word-for-word, faithful = too complex to perceive?, fast, not much controllable speech and sound
   - **Interpreter:** brief, simpler, inter-culture transfer, but how reliable?, slower, controllable
   - We know little what do the target users actually need
4) Risk of no improvement in practice
   - One source always good enough / more sources never good enough.

=> Plan: Exploit the benefits, avoid the risks. Supervise multi-seq to seq model.

Example, where the additional lang. src. helps

ESIC dev 20110215/005_017_EN_Tarand:

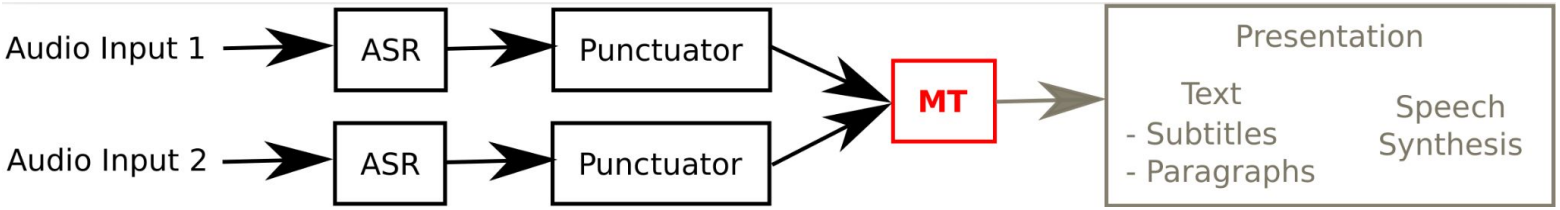
SRC	- Madam President, in my opinion, Mr Werner Schulz has drafted a resolution which is very well **founded** with arguments and draws correct conclusions.
REF	- Paní předsedající, pan Werner Schulz navrhl podle mého názoru usnesení, které je dobře **odůvodněno** argumenty, a jeho výsledkem jsou správné závěry.
En ASR 1	Thank you. In my opinion Mr. Schulz has drafted a resolution which is very well **funded** with arguments and draws correct conclusions.
De SI ASR 1	Herzlichen Dank! Ich denke, dass Herr Schulz eine Entschließung verpasst hat, die wirklich sehr gute Argumente **beinhaltet** und auch die richtigen Schlüsse zieht.
En→Cs	Děkuji vám. Podle mého názoru pan Schulz vypracoval usnesení, které je velmi dobře **financováno** argumenty a vyvozuje správné závěry.
En+De→Cs	Díky. Myslím, že pan Schulz přišel s usnesením, které je velmi dobře **podloženo** argumenty a vyvozuje správné závěry.

Figure 8.5: Cherry-picked example of multi-source translation outperforming single source in the translation of the word "founded." English ASR 1 (Whisper large) following the original speech incorrectly transcribed "funded" instead of "founded." English→Czech single source system translated it wrongly as "financed" ("financováno"), while the multi-source English+German→Czech translated it correctly as "grounded" ("podloženo"). It is very likely thanks to the German ASR 1 (Whisper medium) following German SI that correctly transcribed the corresponding word "beinhaltet" ("contains arguments").

4) Specification

Practical Application

Multi-Source Text-to-text MT in a Cascade



- I focus on the **text-to-text MT part** of cascaded ST
- Primarily for unspecified output modality (speech/text)
- Or for text output, if it matters

Long-Form Monologue

- **Authentic** and challenging
- Often need for simultaneity
- Challenges:
  - **No clear sentence boundaries**
  - Read/spontaneous, fast/slow
  - Hesitations, false starts
  - Language varieties, “accents”
  - Possibly noisy: sounds, other speakers, other languages

Image source: https://www.europarl.europa.eu/

Simultaneous Interpreting (SI)

- **Simultaneous** (vs. consecutive) **interpreting**
- Terms: interpreting ≠ translation

Image source: American Translators Association (ATA)

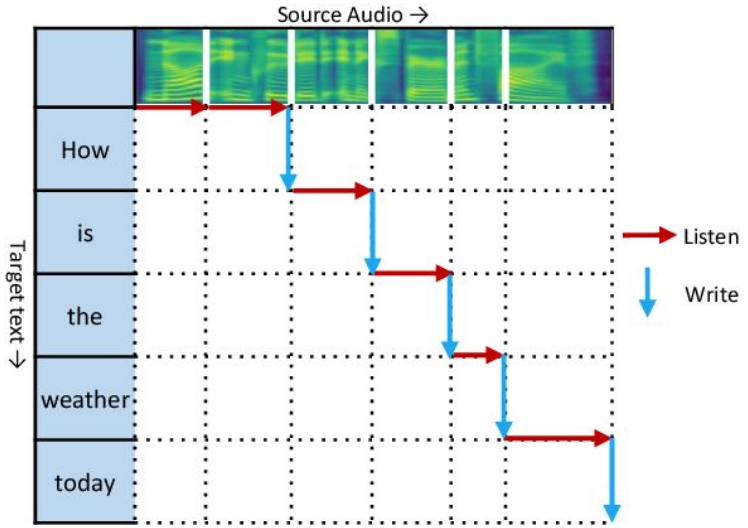
Sim. ST: Re-Translation vs. Streaming

- 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.**

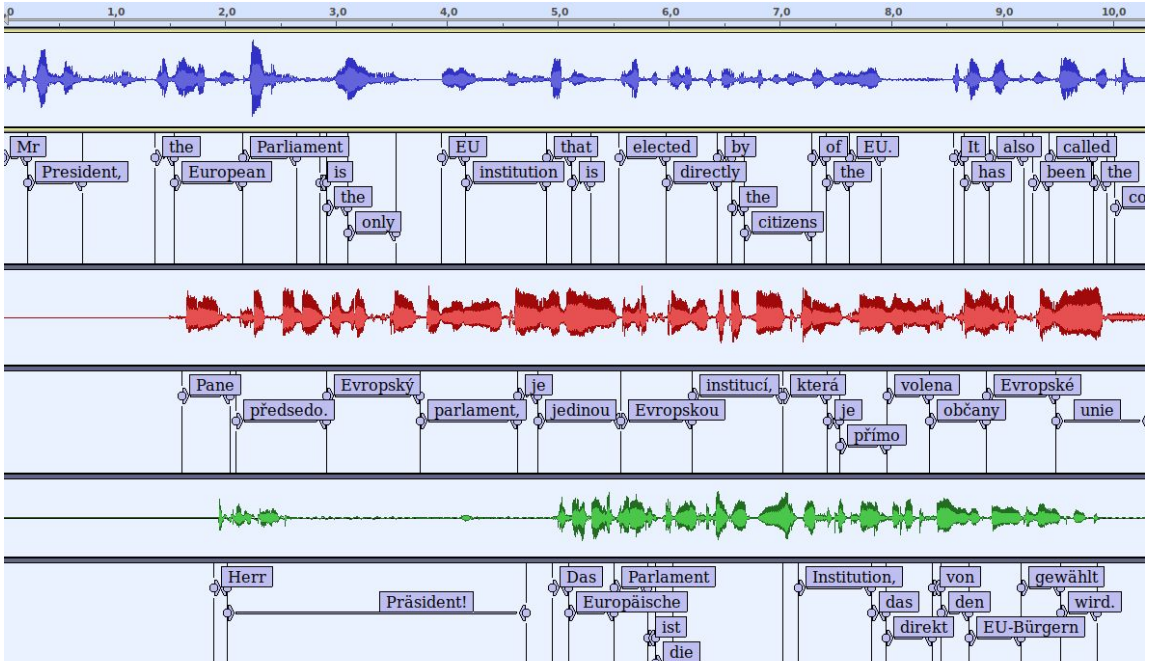


Source: [Ren et al., 2020]

5) Data

ESIC Evaluation Corpus

- Evaluation corpus **ESIC** (Europarl Simultaneous Interpreting Corpus)
  - 3 parallel languages: En + De + Cs
  - En orig + simultaneous interpreting into De, Cs, + parallel revised text translations
  - 10h, 370 speeches, dev+test, word-level timestamps



Ref:
[Macháček et al., 2021]
Lost in Interpreting: Speech Translation from Source or Interpreter?

Interpreting from European Parliament

- EP is a **useful resource** of SI
- 2008-11: translations + SI in 23 langs.!
- Large amount downloaded in 2020, Available for next research
  [Ref: Chapter 4 in thesis]
- Or: VoxPopuli corpus, [Wang et al., 2021]

language tag	language	speeches	speakers	duration	English words
BG	Bulgarian	283	25	7h 16m	70 220
CS	Czech	587	38	15h 55m	167 790
DA	Danish	353	22	8h 39m	99 585
DE	German	3 878	184	110h 44m	1 123 561
EL	Greek	941	56	24h 13m	218 222
EN	English	7 404	396	239h 50m	2 272 520
ES	Spanish	1 302	95	41h 7m	399 147
ET	Estonian	92	7	1h 51m	17 566
FI	Finnish	456	21	9h 33m	94 404
FR	French	2 674	193	100h 55m	1 026 783
GA	Irish	118	9	2h 8m	18 554
HU	Hungarian	841	48	22h 6m	211 309
IT	Italian	2 081	117	53h 39m	495 125
LT	Lithuanian	288	18	5h 43m	54 237
LV	Latvian	124	12	2h 54m	27 098
MT	Maltese	82	6	2h 38m	22 011
NL	Dutch	1 173	60	31h 50m	319 709
PL	Polish	1 787	93	41h 54m	403 246
PT	Portuguese	930	38	25h 52m	235 801
RO	Romanian	1 275	48	27h 31m	245 186
SK	Slovak	777	21	15h 47m	151 648
SL	Slovenian	221	13	5h 16m	50 426
SV	Swedish	576	37	17h 55m	191 228
XM	other/unspec.	242	101	13h 54m	145 367

Table 4.2: Statistics of downloaded and parsed speeches in period 2008/09/01 – 2011/07/04 by language tag in metadata, in alphabetical order. Language tag "XM" stands for other than EU or unspecified language. "English words" stands for the number of words in English translation (or transcript) of the speech.

6) Analysis of Interpreting

Interpreting Analysis (our results)

- Ref: Macháček et al., 2021: Lost in Interpreting: Speech Translation from Source or Interpreter?
- **Shortening**: sim. interpreting is by 13% shorter than offline manual translation
  - En-Cs, average document length in number of syllables, ESIC test
- **Simplification**: words with significantly lower rank in corpus
- **Latency**: intp. 4 sec. behind src, intp.+MT appx. 9.8 sec.
  → similar to relay interpreting, acceptable

Interpreting Strategies (survey)

- (Literature survey, e.g. Čeňková, Ešnerová, Olsen)
- **Segmentation** to sentences: prefer simple sentences, avoid long distance dependencies
  → not 1:1 sentence alignment as in text-to-text translation
- **Language economy**: redundancy reduction (ehm), short variants
- **Generalization**: cats and dogs → pets … short
  a carp → a freshwater fish … when forgot translation
  Hallwang → some village … foreign audience doesn't know it anyway
- **Grammar constructions**: e.g. passivisation in En-Jap. to overcome word-order diff. (He et al., 2016)

Example

- Notice:
  segmentation, shortening, reducing redundancies,
  Hallwang – intercultural transfer + redundant

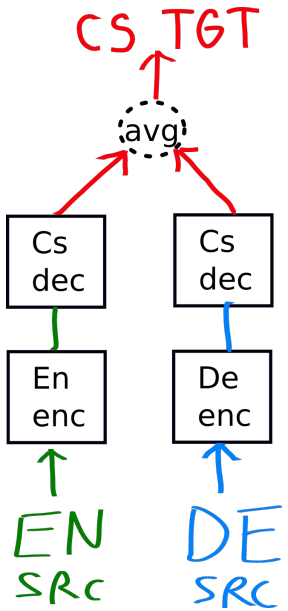
Source (En)	Interpreting (En→Cs)	Gloss to Interpreting
And we try to compare the municipalities with the class of municipalities with the same size,	Zde máme srovnání obcí které mají srovnatelnou velikost.	Here we-have a-comparison of-municipalities, which have a-comparable size.
so we are not comparing Vienna to Hallwang, so we are trying to find similar municipalities **so em** so it will be a fair **compare,** comparison.	Nesrovnáváme tedy nějakou vesnici s Vídní kupříkladu, aby to bylo spravedlivé.	We-are-not-comparing thus some village with Vienna for-instance, so-that it was fair.

Source: Ondřej Bojar at WMT 2020

7) Multi-Sourcing Robustness

Robustness of Multi-Sourcing to Transcription Noise

- Late averaging multi-source NMT [Firat et al., 2016]
  - Like ensembling
  - Text-to-text, sentence-aligned parallel sources – **simplified, unrealistic!**



Robustness of Multi-Sourcing to Transcription Noise

- Mock ASR errors [Martucci et al., 2021]
  - Model edit operations on {gold, ASR transcript} pairs
    => synthethic ASR errors in text test set

0% WER:	Mr President, I would like to thank Mr Brejc for his excellent report.
15% WER:	Present, I would like to thank Mr Brejc for his excellent report.
40% WER:	Makers for President, I would like to thank Me for his report.

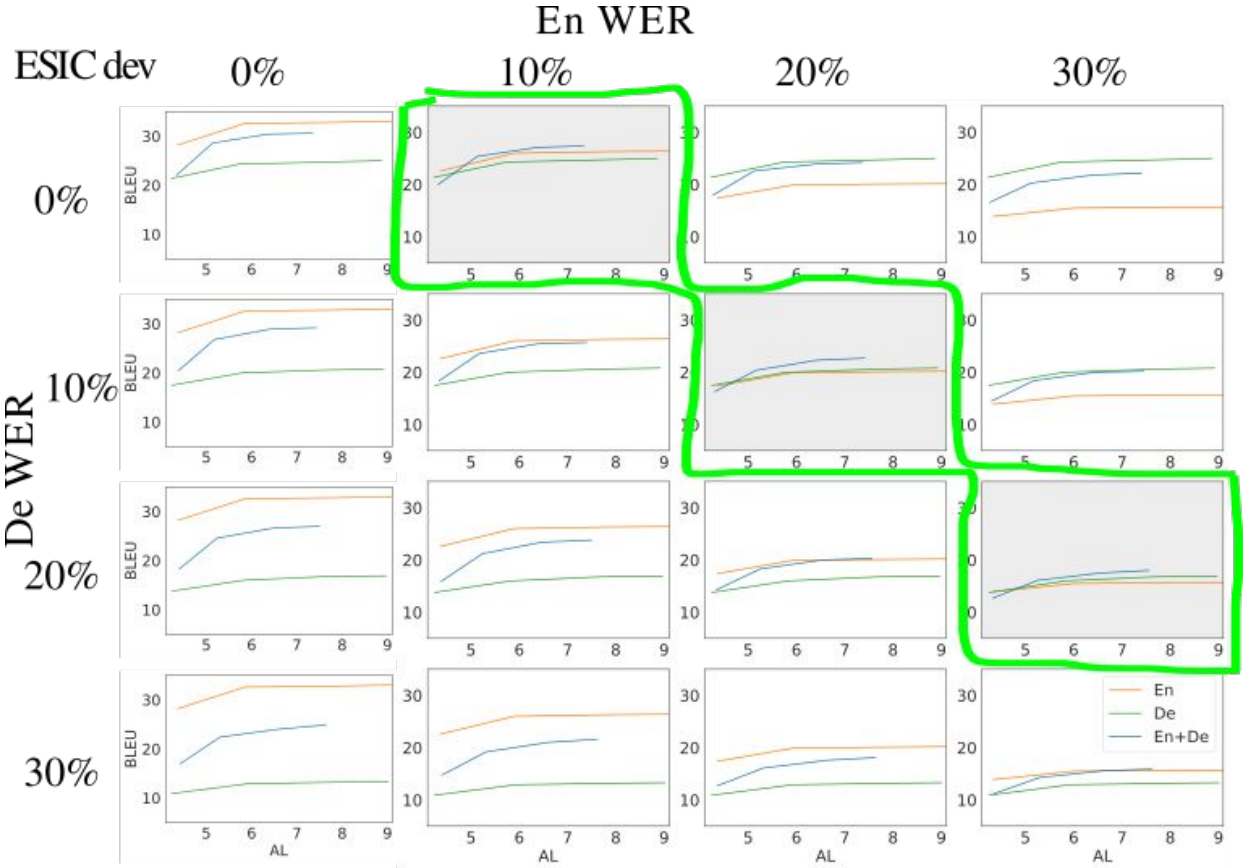
- Simplified setting so far (→ left for future work)
  - mock ASR errors → real ASR
  - text translations → interpretese
  - sentence segmentation → long-form unsegmented speech
  - sentence-aligned sources → not aligned + interpreting delay

Results: Offline with Mock ASR Noise

MULTI-SRC HIGHER

BLEU	ESIC dev single-src.	En WER 0 %	5 %	10 %	15 %	20 %	25 %	30 %	35 %	40 %
s-src.		$33.3^{\pm 0.0}$	$29.7^{\pm 0.3}$	$26.3^{\pm 0.4}$	$22.9^{\pm 0.4}$	$20.4^{\pm 0.5}$	$18.2^{\pm 0.8}$	$15.8^{\pm 0.1}$	$14.0^{\pm 0.2}$	$12.1^{\pm 0.1}$
De WER 0 %	$26.1^{\pm 0.0}$	$31.9^{\pm 0.0}$	$30.0^{\pm 0.2}$	$28.5^{\pm 0.3}$	$26.6^{\pm 0.1}$	$25.2^{\pm 0.4}$	$23.8^{\pm 0.3}$	$21.9^{\pm 0.3}$	$20.5^{\pm 0.2}$	$19.3^{\pm 0.3}$
5 %	$23.5^{\pm 0.0}$	$30.9^{\pm 0.1}$	$29.1^{\pm 0.2}$	$27.6^{\pm 0.3}$	$25.7^{\pm 0.1}$	$24.2^{\pm 0.4}$	$22.8^{\pm 0.4}$	$21.1^{\pm 0.4}$	$19.6^{\pm 0.2}$	$18.6^{\pm 0.2}$
10 %	$21.6^{\pm 0.2}$	$30.0^{\pm 0.2}$	$28.0^{\pm 0.1}$	$26.6^{\pm 0.4}$	$24.6^{\pm 0.3}$	$23.4^{\pm 0.2}$	$21.9^{\pm 0.4}$	$20.2^{\pm 0.1}$	$18.7^{\pm 0.2}$	$17.5^{\pm 0.5}$
15 %	$19.0^{\pm 0.3}$	$28.9^{\pm 0.2}$	$27.1^{\pm 0.1}$	$25.7^{\pm 0.4}$	$23.7^{\pm 0.2}$	$22.4^{\pm 0.4}$	$21.0^{\pm 0.4}$	$19.3^{\pm 0.2}$	$17.8^{\pm 0.3}$	$16.7^{\pm 0.4}$
20 %	$17.1^{\pm 0.3}$	$27.9^{\pm 0.4}$	$26.6^{\pm 0.2}$	$24.9^{\pm 0.4}$	$22.9^{\pm 0.1}$	$21.7^{\pm 0.5}$	$20.0^{\pm 0.4}$	$18.3^{\pm 0.2}$	$17.0^{\pm 0.1}$	$15.7^{\pm 0.1}$
25 %	$15.6^{\pm 0.3}$	$27.1^{\pm 0.3}$	$25.7^{\pm 0.2}$	$24.1^{\pm 0.3}$	$22.1^{\pm 0.2}$	$20.7^{\pm 0.4}$	$19.2^{\pm 0.5}$	$17.4^{\pm 0.2}$	$16.3^{\pm 0.2}$	$14.9^{\pm 0.1}$
30 %	$13.8^{\pm 0.2}$	$25.9^{\pm 0.3}$	$24.5^{\pm 0.4}$	$22.8^{\pm 0.3}$	$20.9^{\pm 0.3}$	$19.6^{\pm 0.2}$	$18.3^{\pm 0.2}$	$16.3^{\pm 0.4}$	$15.1^{\pm 0.1}$	$13.9^{\pm 0.2}$
35 %	$12.5^{\pm 0.2}$	$24.6^{\pm 0.4}$	$22.5^{\pm 0.4}$	$20.9^{\pm 0.2}$	$19.2^{\pm 0.1}$	$18.1^{\pm 0.5}$	$16.7^{\pm 0.3}$	$15.3^{\pm 0.3}$	$14.1^{\pm 0.2}$	$12.9^{\pm 0.1}$
40 %	$10.8^{\pm 0.1}$	$23.4^{\pm 0.4}$	$21.4^{\pm 0.1}$	$20.1^{\pm 0.3}$	$18.3^{\pm 0.5}$	$17.3^{\pm 0.2}$	$16.0^{\pm 0.1}$	$14.4^{\pm 0.1}$	$13.2^{\pm 0.2}$	$12.1^{\pm 0.1}$

Results: Simultaneous with Mock ASR Noise



But: Reference Source Language, Metrics

chrF2	ESIC dev	En WER								
	single-src.	0 %	5 %	10 %	15 %	20 %	25 %	30 %	35 %	40 %
s-src.		$60.2^{\pm0.0}$	$57.2^{\pm0.2}$	$54.4^{\pm0.2}$	$51.4^{\pm0.3}$	$49.2^{\pm0.7}$	$46.8^{\pm0.8}$	$44.3^{\pm0.0}$	$42.1^{\pm0.3}$	$40.1^{\pm0.0}$
0 %	$54.0^{\pm0.0}$	$58.6^{\pm0.0}$	$56.9^{\pm0.2}$	$55.6^{\pm0.1}$	$53.7^{\pm0.2}$	$52.3^{\pm0.5}$	$50.9^{\pm0.4}$	$49.2^{\pm0.3}$	$47.5^{\pm0.2}$	$46.1^{\pm0.2}$
5 %	$51.8^{\pm0.1}$	$57.7^{\pm0.1}$	$56.2^{\pm0.2}$	$54.8^{\pm0.1}$	$52.9^{\pm0.2}$	$51.4^{\pm0.6}$	$50.0^{\pm0.4}$	$48.3^{\pm0.3}$	$46.7^{\pm0.3}$	$45.4^{\pm0.2}$
10 %	$49.9^{\pm0.2}$	$56.8^{\pm0.2}$	$55.1^{\pm0.1}$	$53.7^{\pm0.3}$	$51.8^{\pm0.3}$	$50.4^{\pm0.3}$	$49.0^{\pm0.4}$	$47.3^{\pm0.1}$	$45.6^{\pm0.2}$	$44.3^{\pm0.2}$
15 %	$47.6^{\pm0.3}$	$55.8^{\pm0.0}$	$54.2^{\pm0.1}$	$52.8^{\pm0.3}$	$50.9^{\pm0.2}$	$49.6^{\pm0.4}$	$48.1^{\pm0.5}$	$46.4^{\pm0.3}$	$44.9^{\pm0.3}$	$43.6^{\pm0.1}$
20 %	$45.7^{\pm0.3}$	$54.9^{\pm0.2}$	$53.5^{\pm0.1}$	$51.9^{\pm0.3}$	$50.2^{\pm0.1}$	$48.7^{\pm0.6}$	$47.2^{\pm0.4}$	$45.4^{\pm0.2}$	$43.9^{\pm0.3}$	$42.6^{\pm0.3}$
25 %	$44.0^{\pm0.4}$	$54.2^{\pm0.4}$	$52.9^{\pm0.1}$	$51.3^{\pm0.2}$	$49.3^{\pm0.2}$	$48.1^{\pm0.4}$	$46.5^{\pm0.4}$	$44.7^{\pm0.2}$	$43.3^{\pm0.2}$	$41.7^{\pm0.0}$
30 %	$42.1^{\pm0.3}$	$53.1^{\pm0.3}$	$51.7^{\pm0.3}$	$50.2^{\pm0.2}$	$48.3^{\pm0.3}$	$46.8^{\pm0.3}$	$45.4^{\pm0.5}$	$43.5^{\pm0.3}$	$42.2^{\pm0.2}$	$40.6^{\pm0.1}$
35 %	$40.5^{\pm0.2}$	$52.0^{\pm0.3}$	$50.0^{\pm0.3}$	$48.7^{\pm0.2}$	$46.9^{\pm0.1}$	$45.7^{\pm0.5}$	$44.1^{\pm0.4}$	$42.4^{\pm0.2}$	$41.0^{\pm0.1}$	$39.7^{\pm0.1}$
40 %	$38.6^{\pm0.2}$	$51.1^{\pm0.2}$	$49.2^{\pm0.2}$	$47.8^{\pm0.3}$	$46.0^{\pm0.4}$	$44.8^{\pm0.3}$	$43.2^{\pm0.4}$	$41.5^{\pm0.1}$	$39.8^{\pm0.3}$	$38.6^{\pm0.1}$

De WER

chrF2	news11	En WER								
	single-src.	0 %	5 %	10 %	15 %	20 %	25 %	30 %	35 %	40 %
s-src.		$51.0^{\pm0.0}$	$48.8^{\pm0.2}$	$46.9^{\pm0.1}$	$44.9^{\pm0.1}$	$43.0^{\pm0.1}$	$41.0^{\pm0.1}$	$39.4^{\pm0.1}$	$37.3^{\pm0.1}$	$35.8^{\pm0.0}$
0 %	$50.3^{\pm0.0}$	$50.8^{\pm0.0}$	$49.5^{\pm0.0}$	$48.0^{\pm0.1}$	$46.7^{\pm0.1}$	$45.3^{\pm0.2}$	$43.8^{\pm0.3}$	$42.6^{\pm0.2}$	$41.0^{\pm0.1}$	$39.7^{\pm0.1}$
5 %	$48.3^{\pm0.1}$	$50.0^{\pm0.0}$	$48.6^{\pm0.0}$	$47.3^{\pm0.0}$	$45.7^{\pm0.1}$	$44.5^{\pm0.1}$	$43.1^{\pm0.1}$	$41.8^{\pm0.0}$	$40.1^{\pm0.1}$	$38.8^{\pm0.1}$
10 %	$46.5^{\pm0.2}$	$49.2^{\pm0.1}$	$47.9^{\pm0.2}$	$46.4^{\pm0.1}$	$44.8^{\pm0.1}$	$43.8^{\pm0.0}$	$42.3^{\pm0.0}$	$40.9^{\pm0.1}$	$39.2^{\pm0.2}$	$38.0^{\pm0.1}$
15 %	$44.7^{\pm0.2}$	$48.1^{\pm0.1}$	$46.7^{\pm0.1}$	$45.4^{\pm0.1}$	$43.9^{\pm0.1}$	$42.8^{\pm0.0}$	$41.2^{\pm0.0}$	$40.1^{\pm0.1}$	$38.4^{\pm0.0}$	$37.1^{\pm0.0}$
20 %	$42.9^{\pm0.1}$	$47.1^{\pm0.0}$	$45.8^{\pm0.1}$	$44.3^{\pm0.0}$	$42.9^{\pm0.1}$	$41.7^{\pm0.1}$	$40.4^{\pm0.1}$	$39.1^{\pm0.0}$	$37.4^{\pm0.1}$	$36.3^{\pm0.1}$
25 %	$41.1^{\pm0.1}$	$46.1^{\pm0.2}$	$44.8^{\pm0.0}$	$43.6^{\pm0.1}$	$42.0^{\pm0.1}$	$40.8^{\pm0.1}$	$39.3^{\pm0.2}$	$38.1^{\pm0.1}$	$36.4^{\pm0.1}$	$35.3^{\pm0.1}$
30 %	$39.4^{\pm0.2}$	$45.3^{\pm0.3}$	$43.9^{\pm0.2}$	$42.6^{\pm0.2}$	$41.1^{\pm0.1}$	$39.9^{\pm0.2}$	$38.5^{\pm0.2}$	$37.3^{\pm0.1}$	$35.7^{\pm0.0}$	$34.5^{\pm0.2}$
35 %	$38.0^{\pm0.2}$	$44.3^{\pm0.2}$	$42.9^{\pm0.3}$	$41.5^{\pm0.1}$	$40.2^{\pm0.2}$	$38.9^{\pm0.2}$	$37.6^{\pm0.1}$	$36.4^{\pm0.1}$	$34.9^{\pm0.0}$	$33.7^{\pm0.2}$
40 %	$36.2^{\pm0.2}$	$43.2^{\pm0.2}$	$41.9^{\pm0.2}$	$40.5^{\pm0.2}$	$39.1^{\pm0.1}$	$37.9^{\pm0.1}$	$36.4^{\pm0.2}$	$35.2^{\pm0.1}$	$33.8^{\pm0.2}$	$32.8^{\pm0.2}$

De WER

Table 6.11: chrF2 (avg±stddev) with transcription noise on ESIC dev set whose reference translations were English and on Newstest11 (news11) with balanced reference source language. The area with the green background is where the English single-source outperforms German single-source. Black underlined numbers indicate the area where multi-sourcing achieves higher scores than both single-sourcing options. Red-colored numbers are where at least one single-source scores higher.

Set ref. translation:	Metric	Model En	 De	 De+En
ESIC dev En→Cs	BLEU	*__33.31__	26.13	*31.90
	chrF2	*__60.17__	54.00	*58.59
	En COMET	×__0.920__	0.860	*0.919
	De COMET	×1.007	0.994	*__1.022__
ESIC test En→Cs	BLEU	*__33.63__	27.99	*32.57
	chrF2	*__59.58__	54.75	*58.63
	En COMET	*0.906	0.871	×__0.912__
	De COMET	0.994	×1.006	*__1.018__
news11 3×{De→Cs}	BLEU	16.62 ±0.29	__32.23__ ±0.53	22.47 ±0.44
	chrF2	44.84 ±0.18	__58.81__ ±0.38	49.72 ±0.27
	En COMET	0.528 ±0.002	__0.823__ ±0.002	0.652 ±0.003
	De COMET	0.600 ±0.002	__0.967__ ±0.001	0.757 ±0.003
news11 {De,En,Fr,Es}→Cs, Cs	BLEU	*23.40	22.85	*__23.96__
	chrF2	×__51.00__	50.27	*50.83
	En COMET	0.627	*__0.674__	*0.659
	De COMET	0.700	*__0.832__	*0.766

Same model, real ASR inputs, human eval.

Conclusion:

- multi-sourcing is **able to benefit**, but more issues
- Probably better multi-seq. model, not late averaging

	En+De multi-src	**En single-src**	description
total +	28	**34**	better translated expressions
+	19	28	1 better in segment
++	1	3	2 better in segment
+++	3	0	3 better in segment
–	2	2	worse translated expr. in segment
?	2	3	not graded segment
0	50	43	not +/-/?, or comparable
total	79 segments, 6 documents 1 evaluator, 2 hours		total rated segments

Table 8.2: Results of human evaluation. English single-source system is evaluated as better because it achieved more total + grades (34 vs28) assigned to better-translated words or phrases in the 79 rated segments.

Conclusion

Speech translation from multiple parallel language sources reduces the ASR errors.

Presented paper:

Robustness of Multi-Source MT to Transcription Errors

Macháček et al., Findings ACL 23

Multi-Source SST – References

- *Presented paper:* **Robustness of Multi-Source MT to Transcription Errors**, Macháček, Polák, Bojar, Dabre, Findings ACL 23

  References:

- [Macháček et al., 2021] Lost in Interpreting: Speech Translation from Source or Interpreter?
- [Bojar et al., 2021] Operating a Complex SLT System with Speakers and Human Interpreters
- [Firat et al., 2016] Zero-Resource Translation with Multi-Lingual Neural Machine Translation
- [Martucci et al., 2021] Lexical Modeling of ASR Errors for Robust Speech Translation

8) Simultaneous ST Evaluation

Skip to the slides from IWSLT

Presented paper:

MT Metrics Correlate with Human Ratings of Simultaneous ST

Macháček, Dabre, Bojar, IWSLT 2023

9) Turning Whisper into Real-Time

TL;DR: We made Whisper-Streaming in real-time mode

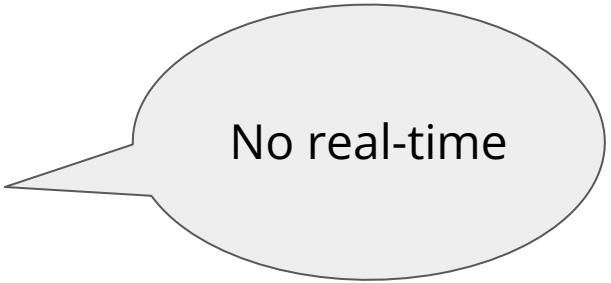
Whisper

speech-to-text

[Radford et al., 2022]

TL;DR: We made Whisper-Streaming in real-time mode

Whisper



speech-to-text

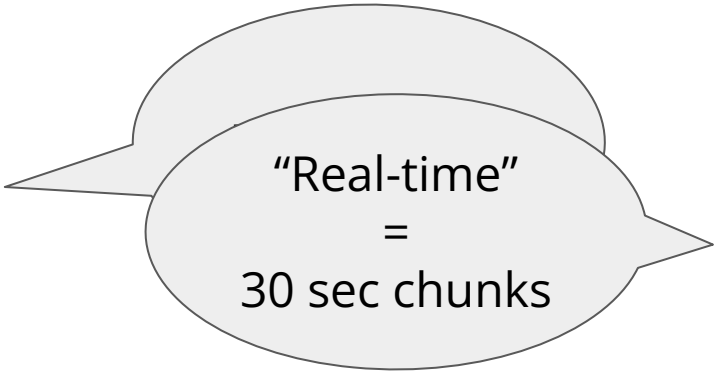
[Radford et al., 2022]

TL;DR: We made Whisper-Streaming in real-time mode

Whisper

speech-to-text

[Radford et al., 2022]



Streaming methods

Local-Agreement

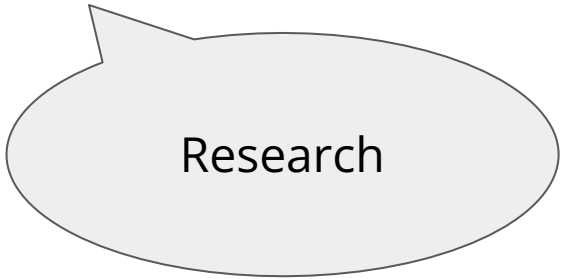
[Liu et al., 2020, Polák et al., 2022, ...]

TL;DR: We made Whisper-Streaming in real-time mode

Streaming methods

Local-Agreement

[Liu et al., 2020, Polák et al., 2022, ...]



TL;DR: We made Whisper-Streaming in real-time mode

Whisper-Streaming

github.com/ufal/whisper_streaming

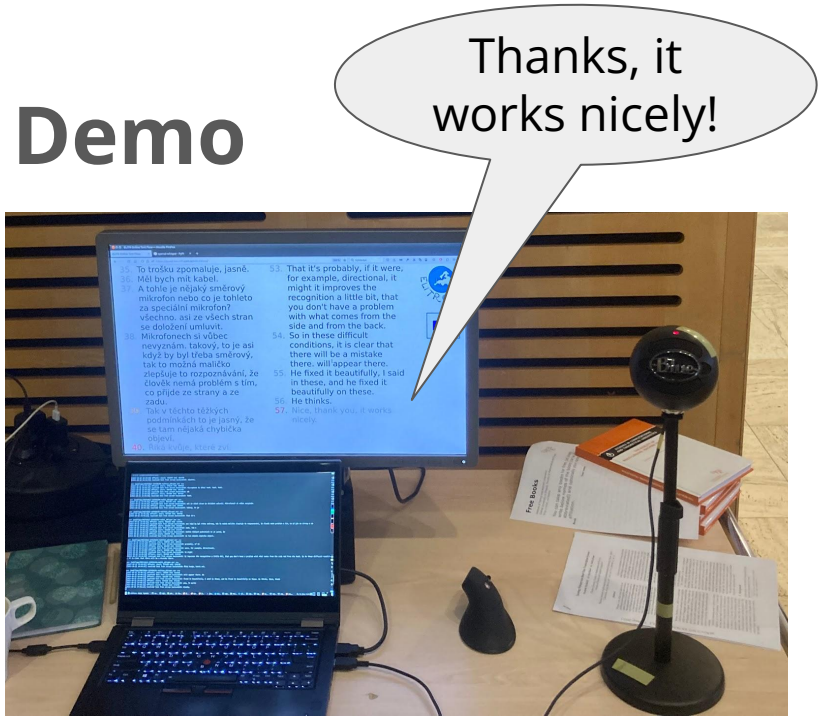
TL;DR: We made Whisper-Streaming in real-time mode

Whisper-Streaming

github.com/ufal/whisper_streaming

Fork 89 | Starred 649

Demo



How it works: Local Agreement-2

With a new audio chunk:

audio buffer

How it works: Local Agreement-2

With a new audio chunk:

1. Append to the audio buffer

How it works: Local Agreement-2

With a new audio chunk:

1. Append to the audio buffer
2. Process buffer -> (VAD)

Voice activity detection

audio buffer

How it works: Local Agreement-2

With a new audio chunk:

1. Append to the audio buffer
2. Process buffer -> (VAD) -> text transcript

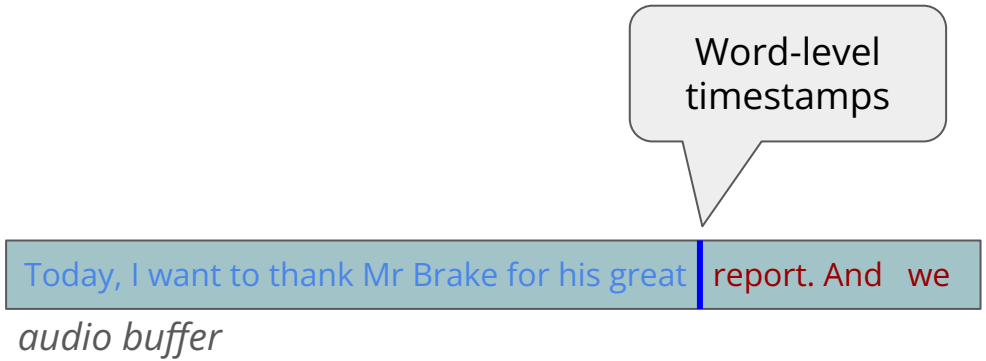
Today, I want to thank Mr Brake for his great report. And we

audio buffer

How it works: Local Agreement-2

With a new audio chunk:

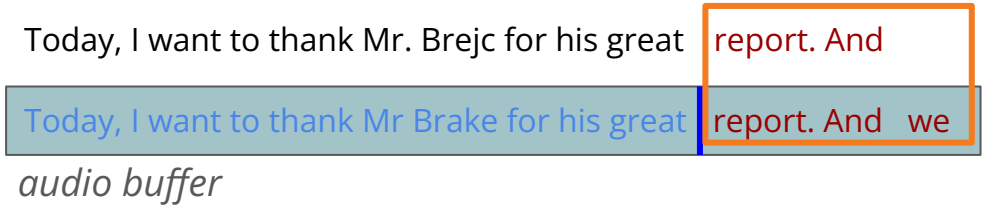
1. Append to the audio buffer
2. Process buffer -> (VAD) -> text transcript
3. Skip previously confirmed part



How it works: Local Agreement-2

With a new audio chunk:

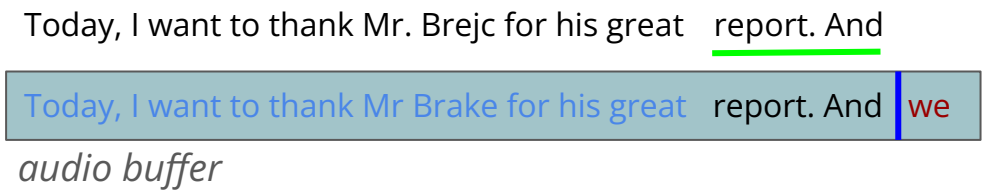
1. Append to the audio buffer
2. Process buffer -> (VAD) -> text transcript
3. Skip previously confirmed part
4. Compare last **2** transcripts



How it works: Local Agreement-2

With a new audio chunk:

1. Append to the audio buffer
2. Process buffer -> (VAD) -> text transcript
3. Skip previously confirmed part
4. Compare last 2 transcripts, confirm common prefix



How it works: Local Agreement-2

With a new audio chunk:

1. Append to the audio buffer
2. Process buffer -> (VAD) -> text transcript
3. Skip previously confirmed part
4. Compare last 2 transcripts, confirm common prefix
5. Trim buffer: on the last segment if buffer > 15s

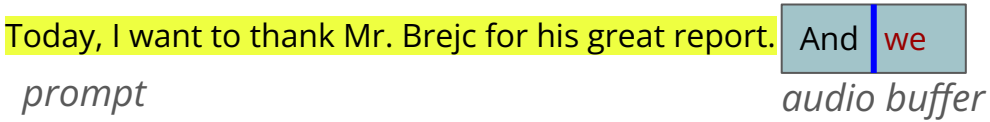
Today, I want to thank Mr. Brejc for his great report. And

audio buffer

How it works: Local Agreement-2

With a new audio chunk:

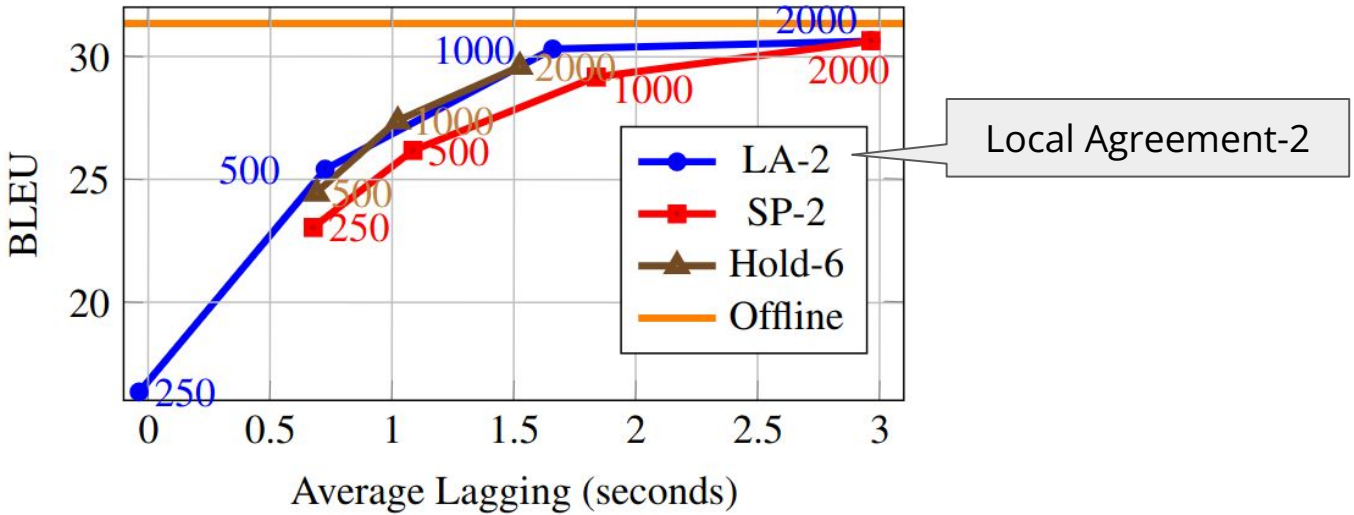
1. Append to the audio buffer
2. Process buffer -> (VAD) -> text transcript
3. Skip previously confirmed part
4. Compare last 2 transcripts, confirm common prefix
5. Trim buffer: on the last segment if buffer > 15s
   update “prompt” = inter-sentence context



Why Local Agreement-2 [Liu et al., 2020]

- Self-adaptive latency = Waits by the uncertainty in language/content
- Best in IWSLT 2022 competition [Polák et al., 2022]
- Min. latency 2-times chunk-size
- Max. unlimited

[CUNI-KIT at IWSLT 22, Polák et al., 2022]



Real-time processing

- Invariant:
  - ~~Buffer starts with a new sentence~~
  - At most 30 seconds

Real-time processing

- Invariant:
  - ~~Buffer starts with a new sentence~~
  - At most 30 seconds

Whisper trained this way

Real-time processing

- Invariant:
  - ~~Buffer starts with a new sentence~~
  - At most 30 seconds

Whisper trained this way

- **faster-whisper** backend on GPU:

Real-time processing

- Invariant:
  - ~~Buffer starts with a new sentence~~
  - At most 30 seconds

Whisper trained this way

- **faster-whisper** backend on GPU:
  - 30 seconds audio ~ in 1 second => real-time

Real-time processing

- Invariant:
  - ~~Buffer starts with a new sentence~~
  - At most 30 seconds

Whisper trained this way

- **faster-whisper** backend on GPU:
  - 30 seconds audio ~ in 1 second => real-time

- Parameter: Update with
  - [MinChunkSize] of new audio,

Real-time processing

- Invariant:
  - ~~Buffer starts with a new sentence~~
  - At most 30 seconds

Whisper trained this way

- **faster-whisper** backend on GPU:
  - 30 seconds audio ~ in 1 second => real-time

- Parameter: Update with
  - [MinChunkSize] of new audio,

0.25 sec/0.5 sec/1.0 sec/ ...

Real-time processing

- Invariant:
  - ~~Buffer starts with a new sentence~~
  - At most 30 seconds

Whisper trained this way

- **faster-whisper** backend on GPU:
  - 30 seconds audio ~ in 1 second => real-time

- Parameter: Update with
  - [MinChunkSize] of new audio, or whatever received.

0.25 sec/0.5 sec/1.0 sec/ ...

Real-time processing

- Invariant:
  - ~~Buffer starts with a new sentence~~
  - At most 30 seconds

Whisper trained this way

- **faster-whisper** backend on GPU:
  - 30 seconds audio ~ in 1 second => real-time

- Parameter: Update with
  - [MinChunkSize] of new audio, or whatever received.

0.25 sec/0.5 sec/1.0 sec/ ...

Processing can take longer.

Real-time processing

- Invariant:
  - ~~Buffer starts with a new sentence~~
  - At most 30 seconds

Whisper trained this way

- **faster-whisper** backend on GPU:
  - 30 seconds audio ~ in 1 second => real-time

- Parameter: Update with
  - [MinChunkSize] of new audio, or whatever received.
  - Controls the latency and quality

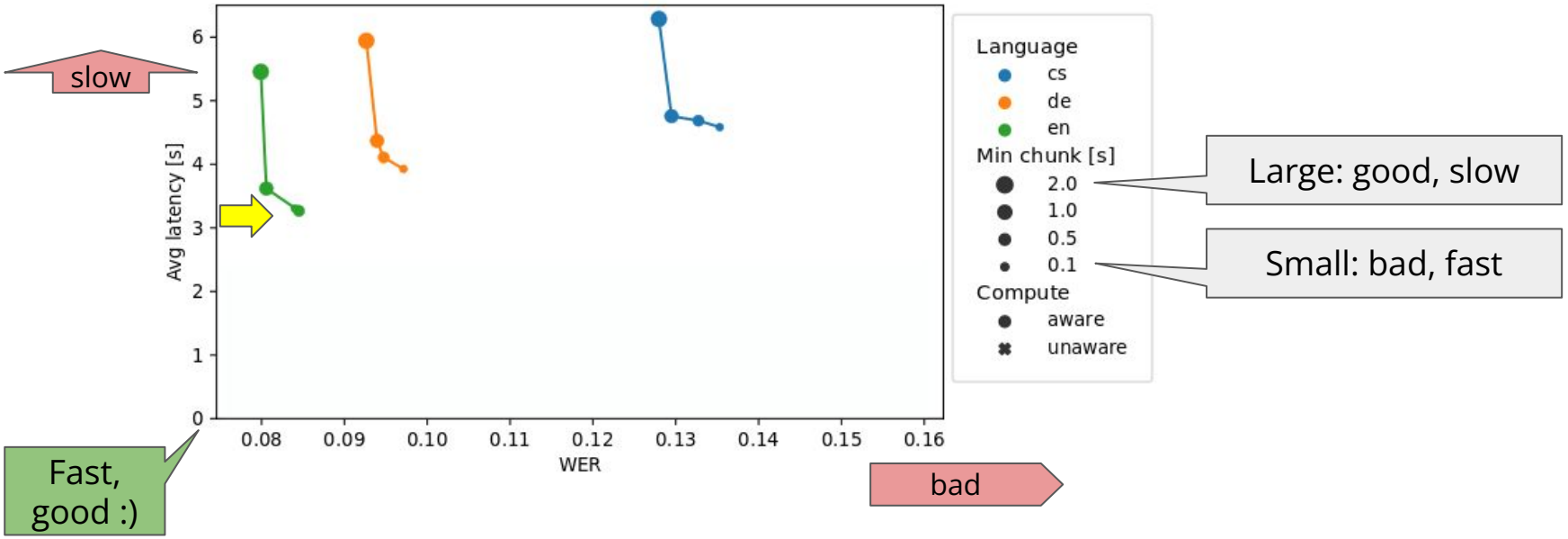
0.25 sec/0.5 sec/1.0 sec/ ...

Processing can take longer.

ASR Performance tests

- ESIC – Europarl, English orig., German, Czech interpreting [Macháček et al., 2021]
- NVIDIA A40 GPU, Whisper large-v2

➡ English 0.5s m.ch.: 8.5% WER, **3.3s** avg. latency

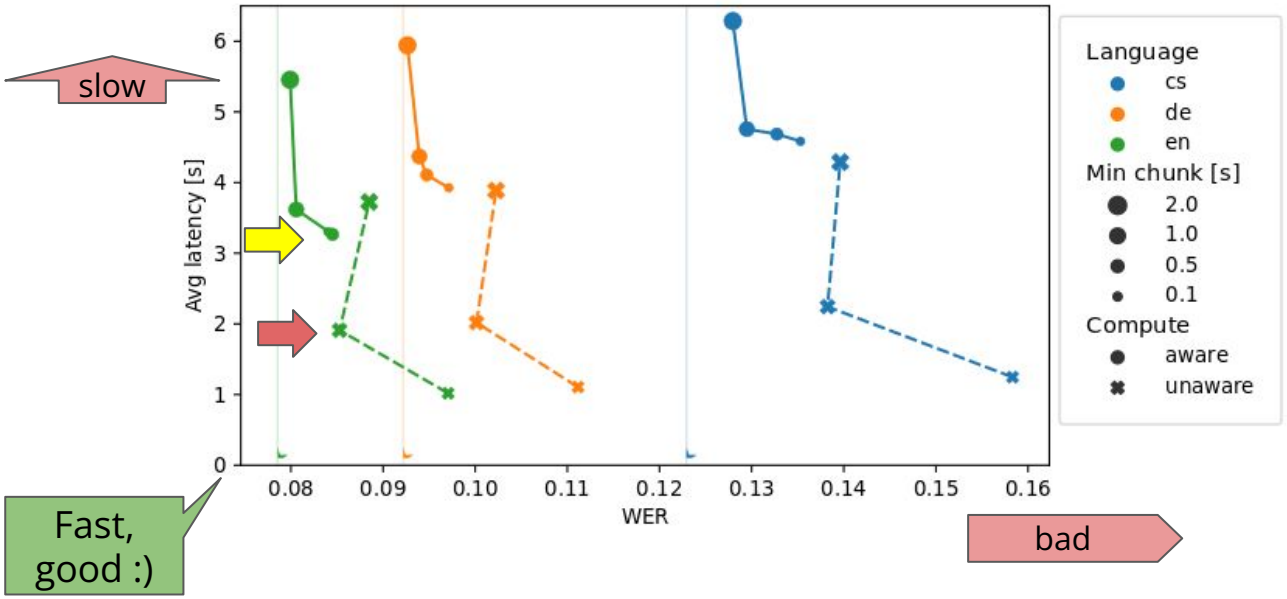


ASR Performance bounds

- Computationally unaware = “optimal hardware speed”
- Offline ASR quality = “optimal quality”
- English 0.5s m.ch.: 8.5% WER, **3.3s** avg. latency -> unaw. +1.2% WER, -2.5s = 1.7s
- Offline: -1.8% WER

Model, language

Implementation, hardware



Demonstration

- Integration with ELITR live speech translation framework
- Evaluation event – one day conference in 🇨🇿, 🇬🇧, 🇺🇦 -> excellent quality 🙂
- **Interactive** demo
  - Speak in any of the **96** langs.! Observe the quality-latency! Have a chat!

New features and bugfixes

- Fix: Trim buffer on segments if buffer > 15s, not on sentences
  - Better quality-latency
  - => no lang. dependent sentence segmenter
  - => no installation issues on Windows and Mac
- Whisper large-v3 model
- **Automatic language detection**
- Seamless Streaming (finished but bad quality)
- Voice activity controller (almost finished)
- OpenAI API (finished)
- Pending – **collaboration welcome** :-)
  - Transcribe and translate at once (batching), multi-clients (batching), last two chunks at once (batching)
  - Forced decoding to skip the processed part of the buffer -> maybe fast enough on CPU?
  - Optimize prompt length
  - Prompting for OOV, ...

Summary

We made Whisper-Streaming in real-time mode

- Speech-to-text with Local-Agreement policy
- Live interactive demo
- Simple and robust implementation
- *Presented paper:*
  Turning Whisper into Real-Time Transcription System,
  Macháček, Dabre, Bojar, IJCNLP-AACL 2023 Demo

`https://github.com/ufal/whisper_streaming`

Whisper-Streaming – References

[Liu et al., 2020] Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis Selection. Proc. Interspeech 2020, 3620-3624, doi: 10.21437/Interspeech.2020-2897

[Macháček et al., 2021] Lost in Interpreting: Speech Translation from Source or Interpreter? Proc. Interspeech 2021, 2376-2380, doi: 10.21437/Interspeech.2021-2232

[Radford et al., 2022] Robust Speech Recognition via Large-Scale Weak Supervision, https://cdn.openai.com/papers/whisper.pdf

[Polák et al., 2022] CUNI-KIT System for Simultaneous Speech Translation Task at IWSLT 2022, In Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 277–285, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

ELITR – European Live Translator, elitr.eu

Dominik Macháček – ufal.cz/dominik-machacek

10) Summary

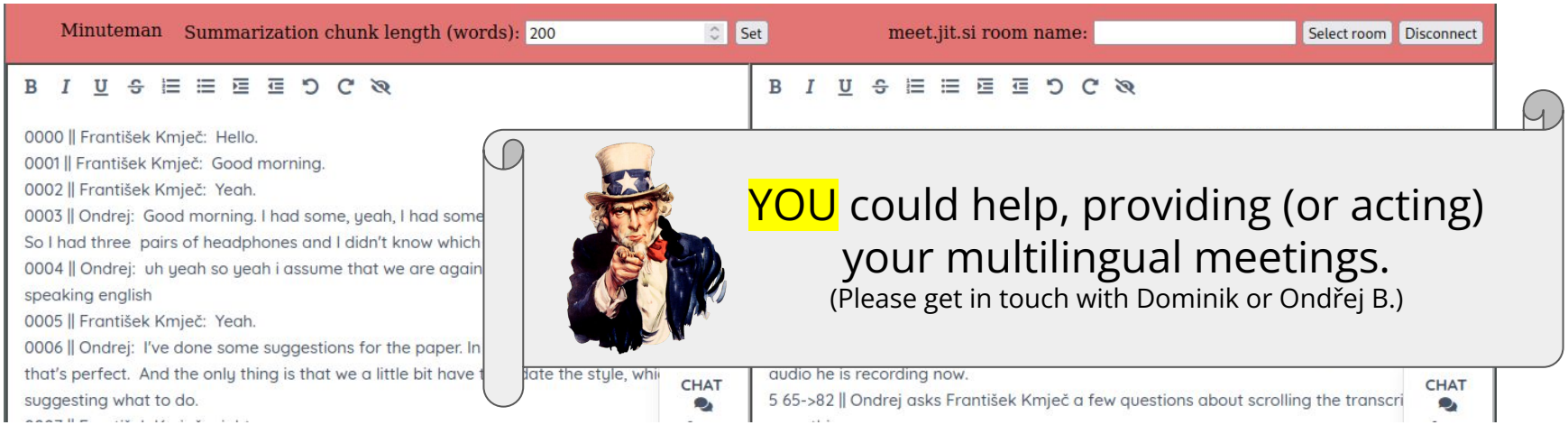
Summary

1. **Conclusion**:
   We investigated methods for multi-source SST from the original and SI
2. **Main Finding:**
   - we set foundations, we experimented with basic model in simplified conditions, we analyzed the weaknesses and proposed further steps.
3. **Main Contributions:**
   a. ESIC
   b. Analysis of SI
   c. Study of multi-sourcing robustness to ASR noise
   d. MT Metrics in SST eval – confirmed previously untested assumption
   e. Whisper-Streaming
4. Further work – InCroMin project
5. Ideas for related work, e.g. prompting in Whisper, live post-editing, …

11) InCroMin

InCroMin: Interactive Crosslingual Minutes

- As a follow-up of ELITR, we got a small funding from the UTTER project.
- Goal:
  - Collect a small corpus of **multilingual meetings**
    - Extreme target: everyone speaks their mother tongue, minutes are created and editable by anyone in any language, everyone gets minutes in their language.
  - Improve Minuteman (https://arxiv.org/pdf/2309.05272.pdf) to support multilinguality.



live transcript, correctable

live minutes, also correctable