# **Multi-Source Simultaneous Speech Translation**

**Dominik Macháček** <u>machacek@ufal.mff.cuni.cz</u> March 4 2024



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



unless otherwise stated

#### 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
- 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

... Chapter 1 ... Section 8.4.5 ... Chapter 2 ... Chapter 3 ... Chapter 4 ... Chapter 5 ... Chapter 6 ... Chapter 7, Section 8.2 ... Section 8.4 ... Chapter 9

#### 1) Long Story Short

#### **Speech Translation**



#### **Simultaneous**

#### **Speech Translation**





#### ...from the <u>Source</u> AND <u>Interpreting</u>



## **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 <u>quest.ms.mff.cuni.cz/elitr/demo/</u>
- 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
  - 80. That was Josef Pazderka from Czech
  - Radio Plus. 81. And now, the President of the Czech
  - Republic, Petr Pavel.
  - Please come up here on stage, and present your opening speech to start the first session of this conference. conference, Ukraine as a Shared Responsibility.
  - 83. Mr. President.
  - 84. Good morning, ladies and gentlemen, guests here and listeners and viewers on the other platforms.
  - 85. When I was asked by the Czech radio to take over the auspices of this event, I did not hesitate for a second, because the topics that we are discussing here today are very important to me.
  - 86. This is the 100th anniversary since the start of the regular broadcast of the Czech radio, which also tells us about the importance of freedom of speech, of talking without censorship, without limitations. the freedom to accept information, to seek information, to spread information, the freedom that in many parts of the world is restricted very strongly, and a freedom... people keep giving their lives for.
  - And specific examples are not far away.
     We have among us the daughter of Boris Nemtsov, the murdered Russian opposition politician, Zhanna Nemtsova.
  - 89. On Vyhorodska street, quite close to the headquarters of the Czech Radio, there is Radio Free Europe, and three of its journalists are now in prison,

- republiky, retra ravia, aby prisei sem k nám a přednesl svůj úvodní projev a vlastně tak otevřel ten první blok celé konference.
- Blok nazvaný Ukrajina jako společná odpovědnost.
- 61. Prosím, pane prezidente.
   62. Dobrý den, dámy a pánové, vážení
- bostý zde v sále, posluchači, ale také diváci na ostatních platformách.
   63. Když mě vedení Českého rozhlasu
- požádalo o záštitu nad dnešní konferenci, nemusel jsem dlouho váhat, protože témata, kterými se tady zabýváme, jsou pro mě velice důležitá.
- 64. Připomínáme si 100. výročí odzahájení pravidelného rozhlasového vysílání a to je zároveň i připomínkou významu svobody slova.
- 65. Svobody vyjadřovat se bez cenzury a bez omezení.
- Svobody přijímat informace a myšlenky, vyhledávat je a křít.
   Svobody, která je v různých koutech světa stále výrazně omezována a za její šprosazování Ildé i dnes platí tu neivvšší cenu.
- 68. Pro konkrétní příklady nemusíme vůbec chodit daleko.
- Mezi námi je dnes dcera zavražděného ruského opozičního politika Borise Němcova, žena Němcovová.
- Na ulici Vinohradská, jen kousek od sídla Českého rozhlasu, sídli i Radio Sobotná Evropa.
- 71. Jehož tři novináři jsou dnes vězněni. -Jihard Losik a Andrej Kuzněčík v Bělorusku a Vladislav Jesipenko na ruském okupovaném Krymu.
- 72. V únoru tohoto roku jsme si připomněli pět let od vraždy slovenského



#### , 83<mark>. is-sur President.</mark>

- Filgħodu tajjeb, nisa u mara, mistiednija hawn u dawk li jisimgħu u l-ispetturi fuq il-
- 85. Meta ntalabni mir-radju Čeka biex tiehu I-awditi ta' dan I-avveniment, ma stajtx ghal sekonda, minhabba li s-suģģetti li qed niddiskutu hawn llum huma importanti hafna ghalija.
- Dan huwa I-100 anniversarju mili-bidu tat-trażmissjoni regolari tar-radju Ček, li jghidilna wkoli dwar I-importanza tallibertà tal-kunsiderazzjoni, ta' tkellem mingħajr ćensura, migħajr limitazzjonijiet. il-libertà li jaćċettaw informazzjoni, li ljfitxu informazzjoni, li jinfirxu informazzjoni, li-libertà li l\*ħafna partijiet tad-dinja hija ristretta ħafna b'saħhtha, u libertà. i.n-n
- U eżempji specifici mhumiex bogħod.
   Aħna għandna fostna t-tifla ta' Boris Nemtsov, il-politika ta' l-oppoźizzjoni Russa maqtula, Zhanna Nemts
- Fuq it-triq ta' Vyhorodska, qrib hafna mill-kwartieri generali tar-Radju Čeka, hemm ir-Radju Ħielsa ta' I-Ewropa, u

вступне слово. г відкрив першии олок 75 конференції. 74 70. Блок під назвою «Україна як спільна

- відповідальність».
- 71. Прошу пана президента.
- Доброго дня, дами та панове, доброго дня гості в залі, слухачі, а
- також глядачі на інших платформах. 73. Коли керівництво Чеського радіо
- попросило мене взяти патронат на цієї конференції, я не вагався. 74. Тому що на теми, про які ми сьогодні
- (4. Іому що на теми, про які ми сьогодні будемо говорити, це теми дуже важливі для мене.
- Сьогодні ми пригадуємо соту річниць від початку трансляції Чеського радіо.
- 1 це також нагадування про важливість свободи слову.
- Свободи висловлювати свою думку без обмежень.
- Свободу приймати інформацію та думки, шукати їх та поширювати.
- Свободу, яка у різних частинах світу досі підається переслідуванням.
   І за неї люди і сьогодні платять
- во. Тза неглюди гсьогодні платя найвищу ціну.
- За такими прикладами нам не треба ходити далеко.
- Сьогодні серед нас є донька вбитого російського політика Жанна Німцова.
- 83. У вулиці Виноградська, зовсім недалеко від місця, де знаходиться Чеське радіо, знаходиться і радіо «Свобода».
- Три журналіста, якого зараз знаходяться за ґратами.
- 85. У лютому цього року

- وثانياً، لم ْنرغْب، بعد هذا أكثر من سنة الصراع، لم .75 نرغب في أن يُنظر إلى هذا كألعاب فيديو. بعض الحركات على الخريطة
- - طوال هذه الألواح، طوال البوم، يجب أن تكون قادراً على رؤية على الأقل نظرة على ذلك. وآمل أن تكون تجربة مثيرة للاهتمام وقوية.
- كَان ذلك (جوزيف بّازدركاً) من الراديوُ التَشيكي . .80 (ب
- - صباح الخبر باً سيداتي وسادة ضيوف هنا والمستمعين والمشاهدين على المنبرات الأخرى وعندما طلبت من الإذاعة التشيكية أن تتولى
- رعابة هذا الحدث، لم أتردد لمدة ثانية، لأن المواضيع التي نناقشها اليوم هامة جدا بالنسبة لى.
- هذاً هو الذكرى السنوية المئوية منذ بداية البت المنتظم للإذاعة التشبكية، التي تُحيرا أيضاً بأهمية حرية العبير، والتحدث بدون رقابة، دون قبود. ونشر المعلومات، والحدث التي تُقدّ في العديد من أنحاء العالم، بفوة جدا، والحرية... الناس وهناك أمثلة محددة ليست عبود، عن ذلك. وهناك أمثلة محددة ليست عبود، عن ذلك. العارمة الروسية المقتلية، إننا بامتسبوق.

الفكارض الروسية المسلم الريب عندام . في شارع فيهورودسكا، قريب جداً من مقر الإذاعة التشيكية، هناك إذاعة أوروبا الحرة، وثلاثة



AŻ

#### 3) Motivation

#### **Benefits and Risks of Source+Interpreter**

1) Quality:

#### Disambiguation

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

#### **Complementary ASR errors:**

 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

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 <b>founded</b> with arguments and draws correct conclusions.
REF	<ul> <li>Paní předsedající, pan Werner Schulz navrhl podle mého názoru usnesení, které je dobře odůvodněno argumenty, a jeho výsledkem</li> </ul>
	jsou správné závěry.
En ASR 1	Thank you. In my opinion Mr. Schulz has drafted a resolution which
	is very well <b>funded</b> 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 <b>beinhaltet</b> und auch die richtigen Schlüsse zieht.
En→Cs	Děkují vám. Podle mého názoru pan Schulz vypracoval usnesení.
2	které je velmi dobře <b>financováno</b> argumenty a vyvozuje správné závěry.
$En+De \rightarrow Cs$	Díky. Myslím, že pan Schulz přišel s usnesením, které je velmi dobře <b>podloženo</b> 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 $\rightarrow$ Czech single source system translated it wrongly as "financed" ("financováno"), while the multi-source English+German $\rightarrow$ 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.**

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



|--|

Source	Outpu	ıt							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	



#### **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



7 0

8.0

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

#### **Interpreting from European Parliament**

•	EP is a <b>useful resource</b> 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,

[<u>Wang et al., 2021</u>]

language tag	language	speeches	speakers	duration	English words
BG	Bulgarian	283	25	7h 16m	70 2 20
CS	Czech	587	38	15h 55m	167790
DA	Danish	353	22	8h 39m	99585
DE	German	3 878	184	110h 44m	1 123 561
EL	Greek	941	56	24h 13m	218 222
EN	English	7404	396	239h 50m	2 272 520
ES	Spanish	1 302	95	41h 7m	399147
ET	Estonian	92	7	1h 51m	17566
FI	Finnish	456	21	9h 33m	94 404
FR	French	2674	193	100h 55m	1 0 2 6 7 8 3
GA	Irish	118	9	2h 8m	18554
HU	Hungarian	841	48	22h 6m	211 309
IT	Italian	2 0 8 1	117	53h 39m	495125
LT	Lithuanian	288	18	5h 43m	54237
LV	Latvian	124	12	2h 54m	27 0 98
MT	Maltese	82	6	2h 38m	22011
NL	Dutch	1 1 7 3	60	31h 50m	319709
PL	Polish	1 787	93	41h 54m	403 246
PT	Portuguese	930	38	25h 52m	235 801
RO	Romanian	1 275	48	27h 31m	245186
SK	Slovak	777	21	15h 47m	151648
SL	Slovenian	221	13	5h 16m	50 4 26
SV	Swedish	576	37	17h 55m	191 228
XM	other/unspec.	242	101	13h 54m	145367

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.
  - $\rightarrow$  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

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

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 ( $\rightarrow$  left for future work)
  - $\circ \quad \text{mock ASR errors} \rightarrow \text{real ASR}$
  - $\circ \quad \text{text translations} \to \text{interpretese}$
  - $\circ$  sentence segmentation  $\rightarrow$  long-form unsegmented speech
  - $\circ$  sentence-aligned sources  $\rightarrow$  not aligned + interpreting delay

#### **Results: Offline with Mock ASR Noise**

#### MULTI-SRC HIGHER

BLEU	ESIC dev	En WER								
	single-src.	0 %	5 %	10 %	/ 15 %	20~%	25 %	30 %	35 %	40 %
s-src		$33.3^{\pm 0.0}$	$29.7^{\pm 0.3}$	$26.3^{\pm 0.4}$	$22.9^{\pm 0.4}$	$20.4^{\pm 0.5}$	$18.2^{\pm 0.8}$	$15.8^{\pm 0.1}$	$14.0^{\pm 0.2}$	$12.1^{\pm 0.1}$
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 %	$21.6^{\pm 0.2}$	$30.0^{\pm 0.2}$	$28.0^{\pm 0.1}$	$26.6^{\pm 0.4}$	$24.6^{\pm 0.3}$	$23.4^{\pm 0.2}$	$21.9^{\pm 0.4}$	$20.2^{\pm 0.1}$	$18.7^{\pm 0.2}$	$17.5^{\pm 0.5}$
H 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 %	$17.1^{\pm 0.3}$	$27.9^{\pm 0.4}$	$26.6^{\pm 0.2}$	$24.9^{\pm 0.4}$	$22.9^{\pm 0.1}$	$21.7^{\pm 0.5}$	$20.0^{\pm 0.4}$	$18.3^{\pm 0.2}$	$17.0^{\pm 0.1}$	$15.7^{\pm 0.1}$
പ്പ് 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**

с	hrF2	ESIC dev		En WER							
		single-src.	0 %	5 %	10 %	15 %	20 %	25 %	30 %	35 %	40 %
	s-src.		$60.2^{\pm 0.0}$	$57.2^{\pm 0.2}$	$54.4^{\pm 0.2}$	$51.4^{\pm 0.3}$	$49.2^{\pm 0.7}$	$46.8^{\pm 0.8}$	$44.3^{\pm 0.0}$	$42.1^{\pm 0.3}$	$40.1^{\pm 0.0}$
	0 %	$54.0^{\pm 0.0}$	$58.6^{\pm 0.0}$	$56.9^{\pm 0.2}$	$55.6^{\pm 0.1}$	$53.7^{\pm 0.2}$	52.3 <sup>±0.5</sup>	$50.9^{\pm0.4}$	49.2 <sup>±0.3</sup>	$47.5^{\pm 0.2}$	46.1 <sup>±0.2</sup>
	5 %	$51.8^{\pm0.1}$	$57.7^{\pm 0.1}$	$56.2^{\pm 0.2}$	$54.8^{\pm 0.1}$	$52.9^{\pm 0.2}$	$51.4^{\pm 0.6}$	$50.0^{\pm0.4}$	$48.3^{\pm0.3}$	$46.7^{\pm 0.3}$	$45.4^{\pm 0.2}$
	10 %	$49.9^{\pm 0.2}$	$56.8^{\pm 0.2}$	$55.1^{\pm 0.1}$	$53.7^{\pm 0.3}$	$51.8^{\pm 0.3}$	$50.4^{\pm 0.3}$	$49.0^{\pm 0.4}$	$47.3^{\pm0.1}$	$45.6^{\pm0.2}$	$44.3^{\pm 0.2}$
ER	15 %	$47.6^{\pm 0.3}$	$55.8^{\pm 0.0}$	$54.2^{\pm 0.1}$	$52.8^{\pm 0.3}$	$50.9^{\pm 0.2}$	$49.6^{\pm 0.4}$	48.1 <sup>±0.5</sup>	$46.4^{\pm0.3}$	44.9 <sup>±0.3</sup>	$43.6^{\pm 0.1}$
M	20 %	$45.7^{\pm 0.3}$	$54.9^{\pm 0.2}$	$53.5^{\pm0.1}$	$51.9^{\pm0.3}$	$50.2^{\pm0.1}$	$48.7^{\pm 0.6}$	$47.2^{\pm0.4}$	$45.4^{\pm0.2}$	$43.9^{\pm 0.3}$	$42.6^{\pm0.3}$
De	25 %	$44.0^{\pm 0.4}$	$54.2^{\pm 0.4}$	$52.9^{\pm 0.1}$	$51.3^{\pm 0.2}$	$49.3^{\pm0.2}$	$48.1^{\pm 0.4}$	$46.5^{\pm0.4}$	$44.7^{\pm 0.2}$	$43.3^{\pm0.2}$	$41.7^{\pm0.0}$
	30 %	$42.1^{\pm 0.3}$	$53.1^{\pm 0.3}$	$51.7^{\pm0.3}$	$50.2^{\pm 0.2}$	$48.3^{\pm0.3}$	$46.8^{\pm 0.3}$	$45.4^{\pm 0.5}$	$43.5^{\pm0.3}$	$42.2^{\pm 0.2}$	$40.6^{\pm 0.1}$
	35 %	$40.5^{\pm 0.2}$	$52.0^{\pm 0.3}$	$50.0^{\pm0.3}$	$48.7^{\pm 0.2}$	$46.9^{\pm0.1}$	$45.7^{\pm 0.5}$	$44.1^{\pm0.4}$	$42.4^{\pm 0.2}$	$41.0^{\pm 0.1}$	$39.7^{\pm0.1}$
	40 %	$38.6^{\pm 0.2}$	$51.1^{\pm 0.2}$	$49.2^{\pm 0.2}$	$47.8^{\pm0.3}$	$46.0^{\pm0.4}$	$44.8^{\pm 0.3}$	$43.2^{\pm0.4}$	$41.5^{\pm0.1}$	$39.8^{\pm0.3}$	$38.6^{\pm0.1}$
с	hrF2	news11				I	En WER	Ē			
с	hrF2	<b>news11</b> single-src.	0 %	5 %	10 %	<b>I</b> 15 %	En WER 20 %	25 %	30 %	35 %	40 %
с 	hrF2 s-src.	news11 single-src.	0%51.0 <sup>±0.0</sup>	$\frac{5\%}{48.8^{\pm 0.2}}$	$\frac{10\%}{46.9^{\pm 0.1}}$	15 % 44.9 <sup>±0.1</sup>	En WER 20 % 43.0 <sup>±0.1</sup>	$\frac{25\%}{41.0^{\pm0.1}}$	30% $39.4^{\pm0.1}$	35% $37.3^{\pm0.1}$	40% $35.8^{\pm0.0}$
с 	hrF2 s-src. 0 %	<b>news11</b> single-src. 50.3 <sup>±0.0</sup>	0% 51.0 <sup>±0.0</sup> 50.8 <sup>±0.0</sup>	5% $48.8^{\pm 0.2}$ $49.5^{\pm 0.0}$	$\frac{10 \%}{46.9^{\pm 0.1}}$ $\frac{48.0^{\pm 0.1}}{2}$	15 % 44.9 <sup>±0.1</sup> 46.7 <sup>±0.1</sup>	En WER 20% $43.0^{\pm 0.1}$ $45.3^{\pm 0.2}$	25% $41.0^{\pm0.1}$ $43.8^{\pm0.3}$	$30 \% \\ 39.4^{\pm 0.1} \\ 42.6^{\pm 0.2}$	35% $37.3^{\pm0.1}$ $41.0^{\pm0.1}$	$\begin{array}{r} 40 \% \\ 35.8^{\pm 0.0} \\ 39.7^{\pm 0.1} \end{array}$
c	hrF2 s-src. 0 % 5 %	news11 single-src. 50.3 <sup>±0.0</sup> 48.3 <sup>±0.1</sup>	$\begin{array}{r} 0 \% \\ 51.0^{\pm 0.0} \\ 50.8^{\pm 0.0} \\ 50.0^{\pm 0.0} \end{array}$	$\frac{5\%}{48.8^{\pm 0.2}}$ $\frac{49.5^{\pm 0.0}}{48.6^{\pm 0.0}}$	$\frac{10\%}{46.9^{\pm 0.1}}$ $\frac{48.0^{\pm 0.1}}{47.3^{\pm 0.0}}$	15%     44.9±0.1     46.7±0.1     45.7±0.1	En WER 20% $43.0^{\pm 0.1}$ $45.3^{\pm 0.2}$ $44.5^{\pm 0.1}$	$\frac{25 \%}{41.0^{\pm 0.1}}$ $\frac{43.8^{\pm 0.3}}{43.1^{\pm 0.1}}$	$\frac{30\%}{39.4^{\pm 0.1}}$ $\frac{42.6^{\pm 0.2}}{41.8^{\pm 0.0}}$	$\begin{array}{r} 35 \% \\ 37.3^{\pm 0.1} \\ 41.0^{\pm 0.1} \\ 40.1^{\pm 0.1} \end{array}$	$\begin{array}{r} 40 \% \\ 35.8^{\pm 0.0} \\ 39.7^{\pm 0.1} \\ 38.8^{\pm 0.1} \end{array}$
с 	hrF2 s-src. 0 % 5 % 10 %	news11 single-src. $50.3^{\pm 0.0}$ $48.3^{\pm 0.1}$ $46.5^{\pm 0.2}$	$\begin{array}{c} 0 \% \\ 51.0^{\pm 0.0} \\ 50.8^{\pm 0.0} \\ 50.0^{\pm 0.0} \\ 49.2^{\pm 0.1} \end{array}$	5% $48.8^{\pm 0.2}$ $49.5^{\pm 0.0}$ $48.6^{\pm 0.0}$ $47.9^{\pm 0.2}$	$\begin{array}{c} 10 \% \\ 46.9^{\pm 0.1} \\ 48.0^{\pm 0.1} \\ 47.3^{\pm 0.0} \\ 46.4^{\pm 0.1} \end{array}$	$\begin{array}{c} 15\%\\ 44.9^{\pm0.1}\\ 46.7^{\pm0.1}\\ 45.7^{\pm0.1}\\ 44.8^{\pm0.1}\end{array}$	En WER 20% $43.0^{\pm 0.1}$ $45.3^{\pm 0.2}$ $44.5^{\pm 0.1}$ $43.8^{\pm 0.0}$	$\begin{array}{c} 25 \% \\ 41.0^{\pm 0.1} \\ 43.8^{\pm 0.3} \\ 43.1^{\pm 0.1} \\ 42.3^{\pm 0.0} \end{array}$	$\begin{array}{r} 30 \% \\ 39.4^{\pm 0.1} \\ 42.6^{\pm 0.2} \\ 41.8^{\pm 0.0} \\ 40.9^{\pm 0.1} \end{array}$	$\begin{array}{r} 35 \% \\ 37.3^{\pm 0.1} \\ 41.0^{\pm 0.1} \\ 40.1^{\pm 0.1} \\ 39.2^{\pm 0.2} \end{array}$	$\begin{array}{r} 40 \% \\ 35.8^{\pm 0.0} \\ 39.7^{\pm 0.1} \\ 38.8^{\pm 0.1} \\ 38.0^{\pm 0.1} \end{array}$
ER   0	hrF2 s-src. 0 % 5 % 10 % 15 %	news11 single-src. $50.3^{\pm 0.0}$ $48.3^{\pm 0.1}$ $46.5^{\pm 0.2}$ $44.7^{\pm 0.2}$	$\begin{array}{c} 0 \% \\ 51.0^{\pm 0.0} \\ 50.8^{\pm 0.0} \\ 50.0^{\pm 0.0} \\ 49.2^{\pm 0.1} \\ 48.1^{\pm 0.1} \end{array}$	$5\% \\ 48.8^{\pm 0.2} \\ 49.5^{\pm 0.0} \\ 48.6^{\pm 0.0} \\ 47.9^{\pm 0.2} \\ 46.7^{\pm 0.1} \\ \end{cases}$	$\begin{array}{c} 10 \% \\ 46.9^{\pm 0.1} \\ 48.0^{\pm 0.1} \\ 47.3^{\pm 0.0} \\ 46.4^{\pm 0.1} \\ 45.4^{\pm 0.1} \end{array}$	$15\%$ $44.9^{\pm0.1}$ $46.7^{\pm0.1}$ $45.7^{\pm0.1}$ $44.8^{\pm0.1}$ $43.9^{\pm0.1}$	En WER 20% $43.0^{\pm0.1}$ $45.3^{\pm0.2}$ $44.5^{\pm0.1}$ $43.8^{\pm0.0}$ $42.8^{\pm0.0}$	$\begin{array}{c} 25 \% \\ 41.0^{\pm 0.1} \\ 43.8^{\pm 0.3} \\ 43.1^{\pm 0.1} \\ 42.3^{\pm 0.0} \\ 41.2^{\pm 0.0} \end{array}$	$\begin{array}{r} 30 \% \\ 39.4^{\pm 0.1} \\ 42.6^{\pm 0.2} \\ 41.8^{\pm 0.0} \\ 40.9^{\pm 0.1} \\ 40.1^{\pm 0.1} \end{array}$	$\begin{array}{c} 35 \% \\ 37.3^{\pm 0.1} \\ 41.0^{\pm 0.1} \\ 40.1^{\pm 0.1} \\ 39.2^{\pm 0.2} \\ 38.4^{\pm 0.0} \end{array}$	$\begin{array}{r} 40 \% \\ 35.8^{\pm 0.0} \\ 39.7^{\pm 0.1} \\ 38.8^{\pm 0.1} \\ 38.0^{\pm 0.1} \\ 37.1^{\pm 0.0} \end{array}$
WER 0	hrF2 s-src. 0 % 5 % 10 % 15 % 20 %	news11 single-src. 50.3 <sup>±0.0</sup> 48.3 <sup>±0.1</sup> 46.5 <sup>±0.2</sup> 44.7 <sup>±0.2</sup> 42.9 <sup>±0.1</sup>	$\begin{array}{c} 0 \% \\ 51.0^{\pm 0.0} \\ 50.8^{\pm 0.0} \\ 50.0^{\pm 0.0} \\ 49.2^{\pm 0.1} \\ 48.1^{\pm 0.1} \\ 47.1^{\pm 0.0} \end{array}$	5% $48.8^{\pm 0.2}$ $49.5^{\pm 0.0}$ $48.6^{\pm 0.0}$ $47.9^{\pm 0.2}$ $46.7^{\pm 0.1}$ $45.8^{\pm 0.1}$	$\begin{array}{c} 10 \% \\ 46.9^{\pm 0.1} \\ 48.0^{\pm 0.1} \\ 47.3^{\pm 0.0} \\ 46.4^{\pm 0.1} \\ 45.4^{\pm 0.1} \\ 44.3^{\pm 0.0} \end{array}$	$15\%$ $44.9^{\pm 0.1}$ $46.7^{\pm 0.1}$ $45.7^{\pm 0.1}$ $44.8^{\pm 0.1}$ $43.9^{\pm 0.1}$ $42.9^{\pm 0.1}$	En WER 20 % $43.0^{\pm 0.1}$ $45.3^{\pm 0.2}$ $44.5^{\pm 0.1}$ $43.8^{\pm 0.0}$ $42.8^{\pm 0.0}$ $41.7^{\pm 0.1}$	$\begin{array}{c} 25 \% \\ 41.0^{\pm 0.1} \\ 43.8^{\pm 0.3} \\ 43.1^{\pm 0.1} \\ 42.3^{\pm 0.0} \\ 41.2^{\pm 0.0} \\ 40.4^{\pm 0.1} \end{array}$	$\begin{array}{c} 30 \% \\ 39.4^{\pm 0.1} \\ 42.6^{\pm 0.2} \\ 41.8^{\pm 0.0} \\ 40.9^{\pm 0.1} \\ 40.1^{\pm 0.1} \\ 39.1^{\pm 0.0} \end{array}$	$\begin{array}{c} 35 \% \\ 37.3^{\pm 0.1} \\ 41.0^{\pm 0.1} \\ 40.1^{\pm 0.1} \\ 39.2^{\pm 0.2} \\ 38.4^{\pm 0.0} \\ 37.4^{\pm 0.1} \end{array}$	$\begin{array}{r} 40 \% \\ 35.8^{\pm 0.0} \\ 39.7^{\pm 0.1} \\ 38.8^{\pm 0.1} \\ 38.0^{\pm 0.1} \\ 37.1^{\pm 0.0} \\ 36.3^{\pm 0.1} \end{array}$
De WER	hrF2 s-src. 0 % 5 % 10 % 15 % 20 % 25 %	news11 single-src. 50.3 <sup>±0.0</sup> 48.3 <sup>±0.1</sup> 46.5 <sup>±0.2</sup> 44.7 <sup>±0.2</sup> 42.9 <sup>±0.1</sup> 41.1 <sup>±0.1</sup>	$\begin{array}{c} 0 \% \\ 51.0^{\pm 0.0} \\ 50.8^{\pm 0.0} \\ 50.0^{\pm 0.0} \\ 49.2^{\pm 0.1} \\ 48.1^{\pm 0.1} \\ 47.1^{\pm 0.0} \\ 46.1^{\pm 0.2} \end{array}$	5% 48.8 <sup>±0.2</sup> 49.5 <sup>±0.0</sup> 48.6 <sup>±0.0</sup> 47.9 <sup>±0.2</sup> 46.7 <sup>±0.1</sup> 45.8 <sup>±0.1</sup> 44.8 <sup>±0.0</sup>	$\begin{array}{c} 10 \% \\ 46.9^{\pm 0.1} \\ 48.0^{\pm 0.1} \\ 47.3^{\pm 0.0} \\ 46.4^{\pm 0.1} \\ 45.4^{\pm 0.1} \\ 44.3^{\pm 0.0} \\ 43.6^{\pm 0.1} \end{array}$	$\begin{array}{c} 15 \% \\ 44.9^{\pm 0.1} \\ 46.7^{\pm 0.1} \\ 45.7^{\pm 0.1} \\ 43.9^{\pm 0.1} \\ 43.9^{\pm 0.1} \\ 42.9^{\pm 0.1} \\ 42.0^{\pm 0.1} \end{array}$	En WER 20 % $43.0^{\pm 0.1}$ $45.3^{\pm 0.2}$ $44.5^{\pm 0.1}$ $43.8^{\pm 0.0}$ $42.8^{\pm 0.0}$ $41.7^{\pm 0.1}$ $40.8^{\pm 0.1}$	$\begin{array}{c} 25 \% \\ 41.0^{\pm 0.1} \\ 43.8^{\pm 0.3} \\ 43.1^{\pm 0.1} \\ 42.3^{\pm 0.0} \\ 41.2^{\pm 0.0} \\ 40.4^{\pm 0.1} \\ 39.3^{\pm 0.2} \end{array}$	$\begin{array}{c} 30 \% \\ 39.4^{\pm 0.1} \\ 42.6^{\pm 0.2} \\ 41.8^{\pm 0.0} \\ 40.9^{\pm 0.1} \\ 40.1^{\pm 0.1} \\ 39.1^{\pm 0.0} \\ 38.1^{\pm 0.1} \end{array}$	$\begin{array}{r} 35 \% \\ 37.3^{\pm 0.1} \\ 41.0^{\pm 0.1} \\ 40.1^{\pm 0.1} \\ 39.2^{\pm 0.2} \\ 38.4^{\pm 0.0} \\ 37.4^{\pm 0.1} \\ 36.4^{\pm 0.1} \end{array}$	$\begin{array}{c} 40 \ \% \\ 35.8^{\pm 0.0} \\ 39.7^{\pm 0.1} \\ 38.8^{\pm 0.1} \\ 38.0^{\pm 0.1} \\ 37.1^{\pm 0.0} \\ 36.3^{\pm 0.1} \\ 35.3^{\pm 0.1} \end{array}$
De WER	hrF2 s-src. 0 % 5 % 10 % 15 % 20 % 25 % 30 %	$\begin{array}{c} \textbf{news11} \\ \textbf{single-src.} \\ \hline 50.3^{\pm 0.0} \\ 48.3^{\pm 0.1} \\ 46.5^{\pm 0.2} \\ 44.7^{\pm 0.2} \\ 42.9^{\pm 0.1} \\ 41.1^{\pm 0.1} \\ 39.4^{\pm 0.2} \end{array}$	$\begin{array}{c} 0 \% \\ 51.0^{\pm 0.0} \\ 50.8^{\pm 0.0} \\ 50.0^{\pm 0.0} \\ 49.2^{\pm 0.1} \\ 48.1^{\pm 0.1} \\ 47.1^{\pm 0.0} \\ 46.1^{\pm 0.2} \\ 45.3^{\pm 0.3} \end{array}$	5% $48.8^{\pm 0.2}$ $49.5^{\pm 0.0}$ $48.6^{\pm 0.0}$ $47.9^{\pm 0.2}$ $46.7^{\pm 0.1}$ $45.8^{\pm 0.1}$ $44.8^{\pm 0.0}$ $43.9^{\pm 0.2}$	$\begin{array}{c} 10 \% \\ 46.9^{\pm 0.1} \\ 48.0^{\pm 0.1} \\ 47.3^{\pm 0.0} \\ 46.4^{\pm 0.1} \\ 45.4^{\pm 0.1} \\ 44.3^{\pm 0.0} \\ 43.6^{\pm 0.1} \\ 42.6^{\pm 0.2} \end{array}$	$15\%$ $44.9^{\pm 0.1}$ $46.7^{\pm 0.1}$ $45.7^{\pm 0.1}$ $43.9^{\pm 0.1}$ $42.9^{\pm 0.1}$ $42.0^{\pm 0.1}$ $41.1^{\pm 0.1}$	En WER 20% $43.0^{\pm 0.1}$ $45.3^{\pm 0.2}$ $44.5^{\pm 0.1}$ $43.8^{\pm 0.0}$ $42.8^{\pm 0.0}$ $41.7^{\pm 0.1}$ $40.8^{\pm 0.1}$ $39.9^{\pm 0.2}$	$\begin{array}{c} 25 \% \\ 41.0^{\pm 0.1} \\ 43.8^{\pm 0.3} \\ 43.1^{\pm 0.1} \\ 42.3^{\pm 0.0} \\ 41.2^{\pm 0.0} \\ 40.4^{\pm 0.1} \\ 39.3^{\pm 0.2} \\ 38.5^{\pm 0.2} \end{array}$	$\begin{array}{c} 30 \% \\ 39.4^{\pm 0.1} \\ 42.6^{\pm 0.2} \\ 41.8^{\pm 0.0} \\ 40.9^{\pm 0.1} \\ 40.1^{\pm 0.1} \\ 39.1^{\pm 0.0} \\ 38.1^{\pm 0.1} \\ 37.3^{\pm 0.1} \end{array}$	$\begin{array}{c} 35 \% \\ 37.3^{\pm 0.1} \\ 41.0^{\pm 0.1} \\ 40.1^{\pm 0.1} \\ 39.2^{\pm 0.2} \\ 38.4^{\pm 0.0} \\ 37.4^{\pm 0.1} \\ 36.4^{\pm 0.1} \\ 35.7^{\pm 0.0} \end{array}$	$\begin{array}{r} 40 \% \\ 35.8^{\pm 0.0} \\ 39.7^{\pm 0.1} \\ 38.8^{\pm 0.1} \\ 38.0^{\pm 0.1} \\ 37.1^{\pm 0.0} \\ 36.3^{\pm 0.1} \\ 35.3^{\pm 0.1} \\ 34.5^{\pm 0.2} \end{array}$
De WER	hrF2 <u>s-src.</u> 0 % 5 % 10 % 15 % 20 % 25 % 30 % 35 %	$news11 \\ single-src. \\ 50.3 \pm 0.0 \\ 48.3 \pm 0.1 \\ 46.5 \pm 0.2 \\ 44.7 \pm 0.2 \\ 42.9 \pm 0.1 \\ 41.1 \pm 0.1 \\ 39.4 \pm 0.2 \\ 38.0 \pm 0.2 \\ \end{array}$	$\begin{array}{c} 0 \% \\ 51.0^{\pm 0.0} \\ 50.8^{\pm 0.0} \\ 50.0^{\pm 0.0} \\ 49.2^{\pm 0.1} \\ 48.1^{\pm 0.1} \\ 47.1^{\pm 0.0} \\ 46.1^{\pm 0.2} \\ 45.3^{\pm 0.3} \\ 44.3^{\pm 0.2} \end{array}$	5% 48.8 <sup>±0.2</sup> 49.5 <sup>±0.0</sup> 48.6 <sup>±0.0</sup> 47.9 <sup>±0.2</sup> 46.7 <sup>±0.1</sup> 45.8 <sup>±0.1</sup> 44.8 <sup>±0.0</sup> 43.9 <sup>±0.2</sup> 42.9 <sup>±0.3</sup>	$\begin{array}{c} 10 \ \% \\ 46.9^{\pm 0.1} \\ 47.3^{\pm 0.0} \\ 46.4^{\pm 0.1} \\ 45.4^{\pm 0.1} \\ 43.6^{\pm 0.1} \\ 42.6^{\pm 0.2} \\ 41.5^{\pm 0.1} \end{array}$	$\begin{array}{c} 15 \% \\ 44.9^{\pm 0.1} \\ 46.7^{\pm 0.1} \\ 45.7^{\pm 0.1} \\ 43.9^{\pm 0.1} \\ 42.9^{\pm 0.1} \\ 42.0^{\pm 0.1} \\ 41.1^{\pm 0.1} \\ 40.2^{\pm 0.2} \end{array}$	En WER 20% $43.0^{\pm0.1}$ $45.3^{\pm0.2}$ $44.5^{\pm0.1}$ $43.8^{\pm0.0}$ $42.8^{\pm0.0}$ $41.7^{\pm0.1}$ $40.8^{\pm0.1}$ $39.9^{\pm0.2}$ $38.9^{\pm0.2}$	$\begin{array}{c} 25 \% \\ 41.0^{\pm 0.1} \\ 43.8^{\pm 0.3} \\ 43.1^{\pm 0.1} \\ 42.3^{\pm 0.0} \\ 41.2^{\pm 0.0} \\ 40.4^{\pm 0.1} \\ 39.3^{\pm 0.2} \\ 38.5^{\pm 0.2} \\ 37.6^{\pm 0.1} \end{array}$	$\begin{array}{c} 30 \ \% \\ 39.4^{\pm 0.1} \\ 42.6^{\pm 0.2} \\ 41.8^{\pm 0.0} \\ 40.9^{\pm 0.1} \\ 39.1^{\pm 0.0} \\ 38.1^{\pm 0.1} \\ 37.3^{\pm 0.1} \\ 36.4^{\pm 0.1} \end{array}$	$\begin{array}{c} 35 \ \% \\ 37.3^{\pm 0.1} \\ 41.0^{\pm 0.1} \\ 40.1^{\pm 0.1} \\ 39.2^{\pm 0.2} \\ 38.4^{\pm 0.0} \\ 37.4^{\pm 0.1} \\ 36.4^{\pm 0.1} \\ 35.7^{\pm 0.0} \\ 34.9^{\pm 0.0} \end{array}$	$\begin{array}{r} 40 \% \\ 35.8^{\pm 0.0} \\ 39.7^{\pm 0.1} \\ 38.8^{\pm 0.1} \\ 38.0^{\pm 0.1} \\ 37.1^{\pm 0.0} \\ 36.3^{\pm 0.1} \\ 35.3^{\pm 0.1} \\ 34.5^{\pm 0.2} \\ 33.7^{\pm 0.2} \end{array}$

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. <u>Black underlined</u> 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	Matuia	Model				
ref. translation:	Metric	En	De	De+En		
	BLEU	*33.31	26.13	*31.90		
ESIC dev	chrF2	*60.17	54.00	*58.59		
En→Cs	En COMET	×0.920	0.860	*0.919		
	De COMET	×1.007	0.994	*1.022		
	BLEU	*33.63	27.99	*32.57		
ESIC test	chrF2	*59.58	54.75	*58.63		
En→Cs	En COMET	*0.906	0.871	×0.912		
	De COMET	0.994	$\begin{array}{c} 26.13\\ 54.00\\ 0.860\\ 0.994\\ 27.99\\ 54.75\\ 0.871\\ ^{\times}1.006\\ \hline \begin{array}{c} 32.23\\ \pm 0.53\\ 58.81\\ \pm 0.38\\ 0.823\\ \pm 0.001\\ \hline \end{array}$	*1.018		
	BLEU	$16.62 \\ \pm 0.29$	32.23 ±0.53	$\underset{\pm 0.44}{22.47}$		
news11	chrF2	$\substack{44.84\\\pm0.18}$	$\underset{\pm 0.38}{\textbf{58.81}}$	$\substack{49.72\\\pm0.27}$		
$3{ imes}{De{ o}Cs}$	En COMET	$\underset{\pm 0.002}{0.528}$	Model           En         De           31         26.13           .17         54.00           120         0.860           007         0.994           63         27.99           58         54.75           006         0.871           094 × 1.006         ±0.53           84         58.81           528         0.823           5002         ±0.002           5002         ±0.002           40.002         ±0.001	$\underset{\pm0.003}{0.652}$		
	De COMET	$\underset{\pm 0.002}{0.600}$	0.967 ±0.001	$\underset{\pm0.003}{0.757}$		
newe11	BLEU	*23.40	22.85	*23.96		
$\{De En Fr Es\} \rightarrow Cs$	chrF2	× 51.00	50.27	*50.83		
Cs	En COMET	0.627	*0.674	*0.659		
25	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 E	n 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 do 1 evaluator, 2 h	cuments 10urs	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.

1) Multi-Source SST

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: <u>Robustness of Multi-Source MT to Transcription</u> <u>Errors</u>, 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

8) SST 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

# Whisper

#### speech-to-text

[Radford et al., 2022]



[Radford et al., 2022]



[Radford et al., 2022]

## **Streaming methods**

Local-Agreement

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

# Streaming methods

Local-Agreement

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



## **Whisper-Streaming**



github.com/ufal/whisper\_streaming



## **Whisper-Streaming**



#### github.com/ufal/whisper\_streaming





With a new audio chunk:

audio buffer

With a new audio chunk:

1. Append to the audio buffer



With a new audio chunk:

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

Voice activity detection

audio buffer

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

With a new audio chunk:

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



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

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

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

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

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

audio buffer

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



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

Today, I want to thank Mr. Brejc for his great report.		we
prompt	audio	buff

#### 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



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

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



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



• **faster-whisper** backend on GPU:

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



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

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



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

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

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



- **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/ ...

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



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

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

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

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



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

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

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

Processing can take longer.

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



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

• Parameter: Update with

 $\bigcirc$ 

 <u>[MinChunkSize]</u> of new audio, or whatever received.

Controls the latency and quality

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

Processing can take longer.

68

#### **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



Implementation, hardware

Model, language

#### Demonstration

- Integration with ELITR live speech translation framework
- Evaluation event one day conference in 🛌, 😹, 💳 -> excellent guality 🙂
- Interactive demo
  - Speak in any of the **96** langs.! Observe the quality-latency! Have a chat! Ο
  - powerrur experience. That was losef Pazderka from Czech Radio Plus.
  - 81. And now, the President of the Czech Republic, Petr Pavel,
  - 82. Please come up here on stage, and present your opening speech to start the first session of this conference. conference, Ukraine as a Shared Responsibility.
  - 83. Mr. President.
  - 84. Good morning, ladies and gentlemen. quests here and listeners and viewers on the other platforms.
  - 85. When I was asked by the Czech radio to take over the auspices of this event, I did not hesitate for a second, because the topics that we are discussing here today are very important to me.
  - 86. This is the 100th anniversary since the start of the regular broadcast of the Czech radio, which also tells us about the importance of freedom of speech. of talking without censorship, without limitations, the freedom to accept information, to seek information, to spread information, the freedom that in many parts of the world is restricted very strongly, and a freedom... people keep giving their lives for.
  - And specific examples are not far away. We have among us the daughter of Boris Nemtsoy, the murdered Russian opposition politician, Zhanna Nemtsova.
  - 89. On Vyhorodska street, guite close to

republiky, relia ravia, aby prisel serii k nám a přednesl svůj úvodní projev a vlastně tak otevřel ten první blok celé konference

- 60. Blok nazvaný Ukrajina jako společná odpovědnost.
- 61. Prosím, pane prezidente.
- 62. Dobrý den, dámy a pánové, vážení hosté zde v sále, posluchači, ale také diváci na ostatních platformách. 63. Když mě vedení Českého rozhlasu požádalo o záštitu nad dnešní konferenci, nemusel isem dlouho váhat.
- protože témata, kterými se tady zabýváme, isou pro mě velice důležitá. Připomínáme si 100. výročí odzahájení
- pravidelného rozhlasového vysílání a to je zároveň i připomínkou významu svobody slova.
- Svobody vyjadřovat se bez cenzury a bez omezení. 66. Svobody přijímat informace a
- myšlenky, vyhledávat je a šířit. Svobody, která je v různých koutech
- světa stále výrazně omezována a za její šprosazování lidé i dnes platí tu nejvyšší cenu.
- 68. Pro konkrétní příklady nemusíme vůbec chodit daleko. 69. Mezi námi je dnes dcera zavražděného
- ruského opozičního politika Borise Němcova, žena Němcovová,
- 70. Na ulici Vinohradská, jen kousek od sídla Českého rozhlasu, sídli i Radio Sobotná Evropa.
- 71. lehož tři novináři isou dnes věznění. -Jihard Losik a Andrej Kuzněčík v Bělorusku a Vladislav Jesipenko na ruském okupovaném Krymu.
- 72. V únoru tohoto roku jsme si připomněli



- 84. Filghodu tajjeb, nisa u mara, mistiednija hawn u dawk li iisimgħu u l-ispetturi fug
- Meta ntalabni mir-radiu Čeka biex tiehu I-awditi ta' dan I-avveniment, ma stajtx għal sekonda, minħabba li s-suġġetti li ged niddiskutu hawn llum huma importanti hafna ghalija
- 86. Dan huwa I-100 anniversariu mill-bidu tat-trażmissioni regolari tar-radiu Ček. li jgħidilna wkoll dwar l-importanza tallibertà tal-kunsiderazzioni, ta' tkellem minghajr censura, minghajr limitazzionijiet. il-libertà li jaccettaw informazzioni. li ifittxu informazzioni. li jinfirxu informazzjoni, il-libertà li f'ħafna partijiet tad-dinia hija ristretta ħafna b'saħħitha, u libertà... in-n
- 87. U eżempji specifici mhumiex bogħod. 88. Ahna ghandna fostna t-tifla ta' Boris Nemtsoy, il-politika ta' l-oppozizzioni
- Russa magtula, Zhanna Nemts 89. Fug it-trig ta' Vyhorodska, grib hafna

вступне слово. Гвідкрив першии олок конференції.

- Блок під назвою «Україна як спільна відповідальність». Прошу пана президента.
- Доброго дня, дами та панове,
- доброго дня гості в залі, слухачі, а також глядачі на інших платформах. 73. Коли керівництво Чеського радіо
- попросило мене взяти патронат на цієї конференції, я не вагався.
- будемо говорити, це теми дуже важливі для мене.
- від початку трансляції Чеського
- 76. І це також нагадування про важливість свободи слову.
- 77. Свободи висловлювати свою думку без обмежень.
- 78. Свободу приймати інформацію та думки, шукати їх та поширювати.
- 79. Свободу, яка у різних частинах світу досі підається переслідуванням.
- 80. І за неї люди і сьогодні платять найвищу ціну.
- 81. За такими прикладами нам не треба ходити далеко.
- 82. Сьогодні серед нас є донька вбитого російського політика Жанна Німцова.
- 83. У вулиці Виноградська, зовсім недалеко від місця, де знаходиться Чеське радіо, знаходиться і радіо «Свобода».
- 84. Три журналіста, якого зараз знаходяться за ґратами. 85. У лютому цього року

وموضوع هذا المؤتمر واضح تماما. للتأكيد على نوعية المُعلّومات التي تأتي إلى الجمهور التشيكي. من المهم ما نوع المعلومات التي نستهلكها.



AŻ

وثانياً، لم نرغّب، بعد هذا أكثر من سنة الصراع، لم .75 نرغب في أن يُنظر إلى هذا كألعاب فيديو. بعض الحركات على الخريطة وما زال هناك مصير فرادي الناس، معاناة فرادي

طوال هذه الألواح، طوال اليوم، يجب أن تكون قادراً على رؤية على الأقل نظرة على ذلك. وآمل أن تكون تجربة مثيرة للاهتمام وقوية. كَان ذلك (جوَريف بَازدركا) من الراديو التشيكي

والآن، رئيس الجمهورية التشيكية، بيتر بابل. رجاءً تعال إلى هنا على المسرح. وعرض خطابك الافتتاحي لبدء الدورة الأولى لهذا المؤتمر. المؤتمر، أوكرانيا كمسؤولية مشتركة. سيدى الرئيس صباح الخير يا سيداني وسادة ضبوف هنا والمستمعين والمشاهدين على المنبرات الأخرى وعندما طلبت من الإذاعة التشيكية أن تتولى رِّعاية هذا الحدث، لم أتردد لمدة ثانية، لأَنَ المواضيع التي نناقشها اليوم هامة جدا بالنسبة

هذا هو الذكري السنوية المئوية منذ بداية البث المنتظم للإذاعة التشبكية، التي تُخبرنا أيضاً بأهمية حربة التعبير، والتحدث بدون رقابة، دون قبود. حرية قبول المعلومات، والبحث عن المعلومات، ونشر المعلومات، والحرية التي تُقيُّد في العديد من أنحاء العالم بقوة جداً، والحرية... الناس يستمرون في إعطاء حياتهم من أجلها. وهناك أُمثلة محددة ليست بعيدة عن ذلك. لدينا بيننا ابنة بوريس نامتسوف، سياسة المعارضة الروسية المقتلة، زاننا نامتسوفا.

89. في شارع فيهورودسكا، قريب جداً من مقر

- 74. Тому що на теми, про які ми сьогодні
- Сьогодні ми пригадуємо соту річниць радіо.

#### **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)
- OpenAl 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
  - $\circ$  Prompting for OOV, ...
#### Summary

# We made <u>Whisper-Streaming</u> 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] <u>Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis</u> <u>Selection</u>. 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] <u>Robust Speech Recognition via Large-Scale Weak Supervision</u>, https://cdn.openai.com/papers/whisper.pdf

[Polák et al., 2022] <u>CUNI-KIT System for Simultaneous Speech Translation Task at IWSLT 2022</u>, 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 – <u>ufal.cz/dominik-machacek</u>



## **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** Ο
    - 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 (<u>https://arxiv.org/pdf/2309.05272.pdf</u>) to support multilinguality. Ο

