FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

ABSTRACT OF DOCTORAL THESIS

Dominik Macháček

Multi-Source Simultaneous Speech Translation

Institute of Formal and Applied Linguistics

Supervisor:	doc. RNDr. Ondřej Bojar, Ph.D.
Study Program:	Computational Linguistics

Prague 2024

The results of this thesis were achieved in the period of a doctoral study at the Faculty of Mathematics and Physics, Charles University in years 2019–2024.

Doctoral Candidate:	Mgr. Dominik Macháček Institute of Formal and Applied Linguistics
Supervisor:	doc. RNDr. Ondřej Bojar, Ph.D. Institute of Formal and Applied Linguistics
Department:	Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University Malostranské náměstí 25, 118 00 Prague 1, Czech Republic
Opponents:	dr. François Yvon ISIR, CNRS & Sorbonne Université Paris, France
	prof. Jan Niehues Karlsruhe Institute of Technology Karlsruhe, Germany
Chairman of Academic Council:	doc. Ing. Zdeněk Žabokrtský, Ph.D. Institute of Formal and Applied Linguistics

This abstract was distributed on May 23, 2024.

The thesis defence will take place on June 6, 2024 at 10:00 a.m. in front of a committee for thesis defences in the branch Mathematical Linguistics at the Faculty of Mathematics and Physics, Charles University, Malostranské nám. 25, Prague 1, room S1.

The thesis can be viewed at the Study Department of Doctoral Studies of the Faculty of Mathematics and Physics, Charles University, Ke Karlovu 3, Prague 2.

MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

AUTOREFERÁT DISERTAČNÍ PRÁCE

Dominik Macháček

Simultánní překlad řeči z více zdrojů

Ústav formální a aplikované lingvistiky

Školitel:	doc. RNDr. Ondřej Bojar, Ph.D.
Studijní program:	Matematická lingvistika

Praha 2024

Disertační práce byla vypracována na základě výsledků získaných během doktorského studia na Matematicko-fyzikální fakultě Univerzity Karlovy v letech 2019–2024.

Doktorand: Mgr. Dominik Macháček
Ústav formální a aplikované lingvistiky

Školitel: doc. RNDr. Ondřej Bojar, Ph.D.
Ústav formální a aplikované lingvistiky

Školicí pracoviště: Ústav formální a aplikované lingvistiky,
Matematicko-fyzikální fakulta,
Univerzita Karlova
Malostranské náměstí 25,
118 00 Praha 1, Česká republika

Oponenti: dr. François Yvon
ISIR, CNRS & Sorbonne Université
Paříž, Francie

prof. Jan Niehues
Karlsruhe Institute of Technology
Karlsruhe, Německo

Předseda RDSO: doc. Ing. Zdeněk Žabokrtský, Ph.D.
Ústav formální a aplikované lingvistiky

Autoreferát byl rozeslán dne 23. května 2024.

Obhajoba disertační práce se koná dne 6. června 2024 v 10:00 před komisí pro obhajoby disertačních prací v oboru Matematická lingvistika na Matematicko-fyzikální fakultě UK, Malostranské nám. 25, Praha 1, v místnosti S1.

S disertační prací je možno se seznámit na studijním oddělení Matematicko-fyzikální fakulty UK, Ke Karlovu 3, Praha 2.

1. Introduction

Simultaneous speech translation is a technology that aims to assist with understanding foreign language simultaneously, to enable real-time interaction between the speaker and the audience. Nowadays, it often faces issues with quality, especially when the source speech is noisy, non-standard, or contains specific terminology.

In many cases, e.g. in international multi-lingual organizations such as United Nations (UN), European Union (EU), or in European Organisation of Supreme Audit Institutions (EUROSAI), there is human simultaneous interpreting of the source speech into multiple official languages. In our thesis, we investigate the opportunity to leverage multiple speech signals in parallel languages, the original and one or more simultaneous interpreting, as sources of simultaneous speech translation. The expected benefits are quality gains due to word sense disambiguation and higher robustness to transcription errors.

Focus We primarily focus on the text-to-text machine translation (MT) component of cascaded simultaneous speech translation (SST). We focus on the setup that can be used in a practical application as in Figure 1.1. We primarily focus on the long-form monologue speech, which means a continuous speech of multiple utterances in a row, without explicit sentence boundaries.

Original Plan We expected to leverage and combine state-of-the-art solutions for the multi-source text-to-text neural machine translation (NMT) – multi-sequence to sequence learning as e.g. in Zoph and Knight (2016); Firat et al. (2016); Dabre et al. (2017), and the state-of-the-art SST methods – streaming policies applied to the offline model (Papi et al., 2023a,b; Polák et al., 2023). Furthermore, we planned to synthesize the original and simultaneous interpreting training data, using parallel text data and time offsets learned on real simultaneous interpreting data and either use authentic interpreting data from the European Parliament, or synthesize also the interpreting style data.

Main Contributions Our results after four years of PhD.:

1. ESIC – an evaluation corpus for multi-source SST with simultaneous interpreting from the European Parliament (Chapter 4, Macháček et al., 2021).

2. Analyses of simultaneous interpreting as a source of SST (Chapter 5, Macháček et al., 2021).

3. Experiments showing that multi-sourcing leads to robustness to ASR errors (Chapter 6, Macháček et al., 2023c).

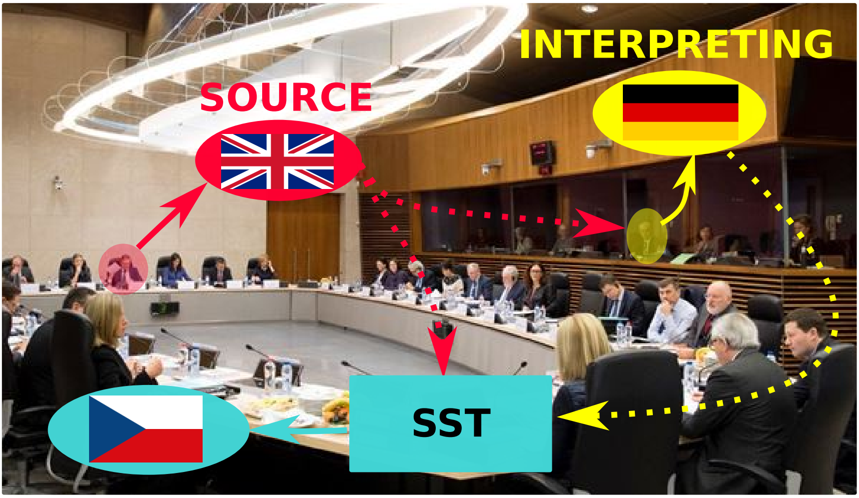


Figure 1.1: Illustration of the use case and setup of multi-source simultaneous speech translation. There is the source speaker using English, parallel simultaneous interpreting from English to German, and SST (machine) combining the English and German sources for translating into the Czech target.

4. Evidence that the offline text-to-text MT metrics are reliable in simultaneous mode (Chapter 7, Macháček et al., 2023a).
5. Whisper-Streaming – a practical tool that makes large offline ASR model work in simultaneous mode (Chapter 8, Macháček et al., 2023b).
6. We thoroughly describe the motivation, challenges, considered options, state of the art, and our experiments with details for reproductions (in this whole thesis, Chapters 1-9) to inspire others as much as possible.

Main Finding Multi-source simultaneous speech translation from the original speech and parallel simultaneous interpreting may bring quality gains in certain situations, especially when the speech recognition quality of the sources is similar (Chapter 6 and Chapter 8, Macháček et al., 2023c).

2. Motivation

Let us summarize what leads us to an investigation of multi-source simultaneous speech translation: the expected benefits and the risks and research challenges. Furthermore, in Chapter 2, we also describe the most related work, showing that the task had not been thoroughly investigated but ready to work on.

2.1 Benefits

The expected benefits of using multiple speech signals in simultaneous speech translation (SST), the original and parallel simultaneous interpreting into auxiliary language, are **gains in translation quality**. They come from the two sources:

1. **Word-sense disambiguation.** As Firat et al. (2016) and others showed on the text-to-text neural machine translation (NMT), multiple parallel language sources can help to enhance the translation quality. It is assumed that it is due to the word-sense disambiguation complemented by the other language. For example, the English word *alien* is ambiguous, it can mean the foreigner (German *Fremde*) or extraterrestrial (German *Außerirdisher*). A parallel German word disambiguates it. And vice-versa, an ambiguous German word *Schloss* can be either English *lock* or *castle*.

2. **Complementary speech recognition errors.** In speech translation, there are often errors stemming from the incorrect recognition of the source word (Ruiz and Federico, 2014; Ruiz et al., 2017; Xue et al., 2020; Martucci et al., 2021), e.g. due to homonymy (*ate* vs. *eight*), noise, non-standard speech, or learning error. Parallel language sources can complement each other, as we illustrate in Figure 2.1.

Another example where multi-sourcing helps the translation quality is in Figure 2.2.

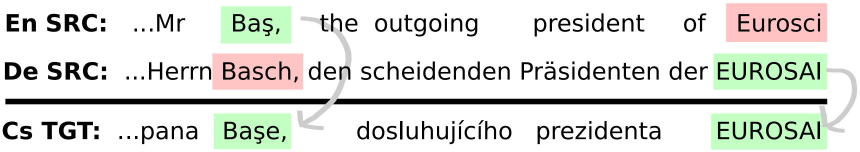

En SRC: ...Mr Baş, the outgoing president of Eurosci
De SRC: ...Herrn Basch, den scheidenden Präsidenten der EUROSAI
Cs TGT: ...pana Başe, dosluhujícího prezidenta EUROSAI

Figure 2.1: Example of how complementary speech recognition errors from two parallel language sources (English and German) can be used to benefit the target translation (into Czech).

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	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	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 2.2: Cherry-picked example of multi-source translation outperforming single source in the translation of the word “founded.” English ASR (model 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 (model Whisper medium) following German simultaneous interpreting (SI) that correctly transcribed the corresponding word “beinhaltet” (“contains arguments“).

The second expected benefit of multi-sourcing in SST is that there has to be **no human intervention needed** for detecting and switching the optimal source from multiple options, from the original and multiple simultaneous interpreting. For humans, it is very risky and challenging to follow and switch between multiple streams at once, in real-time. For example, at the EUROSAI Congress that was operated by the ELITR project (Bojar et al., 2021b), English was available in 4 parallel variants.

The third expected benefit of multi-sourcing from the original and simultaneous interpreting is the possibility to **leverage the advantages of both** options, of the direct translation of the source, and translating the simultaneous interpreter. The direct translation tends to be fast, while simultaneous interpreting brings additional delay. The direct translation may be more accurate and more word-for-word than interpreting. On the other hand, it may be a disadvantage: when the original is noisy, too verbose, or not fluent, it may be too difficult to follow for the target users. Simultaneous interpreters condense the original message, reduce redundancies, and simplify, which may be useful. It is also possible to use the inter-cultural transfer provided by simultaneous interpreting.

2.2 Risks

Latency In terms of translation latency, interpreting creates a delay. Translation from interpreting is more delayed than from the original, however, in Macháček et al. (2021) and in Chapter 5 we measured that the delay is acceptable.

Risk of no improvement in practice There may be no room where multi-sourcing may be beneficial, between one source being too much helpful (e.g. high ASR quality) and all the sources unusable at all (e.g. too low ASR quality). We investigate this risk in Chapter 6.

Challenges Multi-source SST consists of several challenging subtasks, e.g. simultaneous low-latency speech recognition, aligning the sources – original and interpreting, and segmentation to translation units because simultaneous interpreting usually does not translate one source sentence into one target sentence, in contrast to text-to-text translation.

The next challenge is to obtain training and evaluation data to train NMT for multi-sourcing with the original and interpreting. Another challenge is to make NMT working in the simultaneous mode, to translate incomplete sentence prefixes right at the time when the speaker is uttering them, in low, real-time latency.

Next, there are challenges of the speech source modality: ambiguity at all language layers from phonetics to syntax, noise, speaker adaptation, etc. Handling speech recognition errors is challenging.

Last, but not least challenge of multi-sourcing is to reconstruct the original meaning from the multiple noisy sources. The options are voting when we have at least three sources, confidence scores from the sources, or a neural network to detect the confidence by itself, from supervision and sufficient context.

Risk of high complexity Since the subtasks are in a sequence, they influence each other. It is possible that we substantially advance performance in our main subtask, but the entire solution fails because of low performance in another subtask beyond our scope. However, there is a simple mitigation strategy: do research in small subsequent steps, focus only on some subtasks, and use existing or not yet existing but reasonably expected solutions for the other subtasks.

3. Results

Evaluation Data To measure the progress of our research, we needed evaluation data. In Chapter 4, we summarize potentially usable evaluation and training data, the reasons why we created a new evaluation dataset ESIC, and the semi-automatic creation process, from downloading from the web of the European Parliament to the selection of 10 hours for human annotation. We also describe how we aligned multi-parallel sentences and segments.

ESIC (Europarl Simultaneous Interpreting Corpus, Macháček et al., 2021) contains 10 hours of authentic English speeches from the European Parliament from the period 2008-2011, and parallel simultaneous interpreting into German and Czech with human transcripts of different verbosity levels (including or excluding false starts, interruptions, unintelligibles, etc.), with word-level timestamps and with parallel revised text translations. A preview of the corpus is in Figure 3.1. The corpus is available at `http://hdl.handle.net/11234/1-5415`.

Interpreting Analysis In Chapter 5 and in Macháček et al. (2021), we investigate potential challenges of simultaneous interpreting used as an auxiliary source in multi-source simultaneous speech translation. We measure and consider the latency of interpreting and of speech translation following the interpreting, and we realize that it is feasible for the simultaneous use case. We also survey literature for the issues of quality and interpreting strategies which often make interpreting not directly parallel to the original.

Robustness to ASR Errors Then, in Chapter 6 and in Macháček et al. (2023c) we present the main research result of our work. We study the robustness of multi-sourcing to transcription errors.

First, we found a statistically significant evidence that the ASR errors in two parallel language sources, the original and simultaneous interpreting, are independent, which means that they can complement each other.

Then, we create a multi-sourcing model for simplified conditions: for parallel, optimally sentence-aligned text translations. We also disregard the time offset of simultaneous interpreting. We use late averaging, a multi-source model by Firat et al. (2016) (see Figure 3.2).

Then, we evaluate the multi-source model on various levels of ASR errors. Since we do not have many ASR models with various error levels, we synthesize the ASR errors in the text inputs. We model them using the unigram noise model by Martucci et al. (2021) which we extend to produce any target level of ASR errors. The model learns the edit operations on the pairs of gold and ASR transcripts, so it can e.g. realistically shuffle similarly sounding words. We show an example in Figure 3.3.

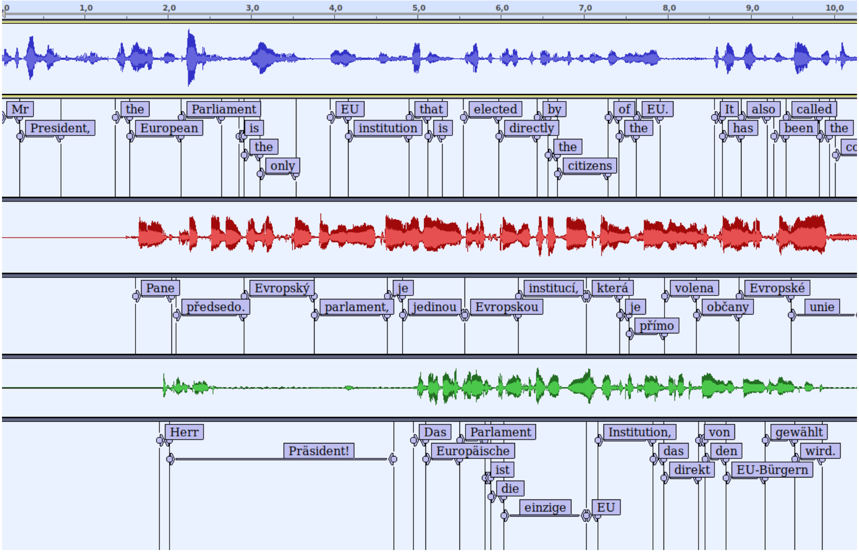


Figure 3.1: Preview of word-level timestamps attached to audio in ESIC. We can see the original English audio track at the top (depicted as a blue oscillogram), followed by words of the manual transcripts (in light violet rectangles) that are automatically aligned to the time segments (black vertical lines with round "knobs") when the word was uttered in the audio. The rectangles sometimes overlap the segments, but it is only a matter of visualization. In the middle and at the bottom, there are parallel audio and timestamped transcript tracks of Czech and German simultaneous interpreting. They start with several seconds of silence when the interpreters either wait for a meaningful translation unit or are occupied by the previous speech.

The results (Table 3.1) indicate that multisourcing may bring quality gains, especially when the source ASRs are of comparable quality. However, we also noticed that the measured gains are smaller with other metric than BLEU and that the source language from which the reference was translated has an effect.

Last, but not least, we implemented simultaneous streaming mode to the multisourcing model using LocalAgreement (Polák et al., 2022), the state-of-the-art simultaneous policy. The results in the simultaneous mode are analogical to those in the offline mode.

Evaluation Questions In Chapter 7 and in Macháček et al. (2023a), we focused on practical questions regarding the evaluation of simultaneous speech translation.

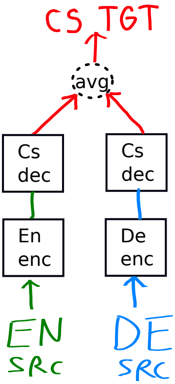


Figure 3.2: Late averaging multi-source model.

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.

Figure 3.3: Example of synthetic automatic speech recognition (ASR) errors in the clean text sources. The first line, 0% WER, is the correct, gold transcript. In the second line, there are two errors, deletion of "Mr" and substitution of "President" to "Present." There are 13 words in the gold transcript, the WER is therefore 2/13 = 15%.

	BLEU	**ESIC dev**	**En WER**								
		single-src.	0 %	5 %	10 %	15 %	20 %	25 %	30 %	35 %	40 %
	s-src.		$33.3^{\pm 0.0}$	$29.7^{\pm 0.3}$	$26.3^{\pm 0.4}$	$22.9^{\pm 0.4}$	$20.4^{\pm 0.5}$	$18.2^{\pm 0.8}$	$15.8^{\pm 0.1}$	$14.0^{\pm 0.2}$	$12.1^{\pm 0.1}$
De WER	0 %	$26.1^{\pm 0.0}$	$31.9^{\pm 0.0}$	$30.0^{\pm 0.2}$	$\mathbf{28.5}^{\pm \mathbf{0.3}}$	$26.6^{\pm 0.1}$	$25.2^{\pm 0.4}$	$23.8^{\pm 0.3}$	$21.9^{\pm 0.3}$	$20.5^{\pm 0.2}$	$19.3^{\pm 0.3}$
	5 %	$23.5^{\pm 0.0}$	$30.9^{\pm 0.1}$	$29.1^{\pm 0.2}$	$27.6^{\pm 0.3}$	$\mathbf{25.7}^{\pm \mathbf{0.1}}$	$24.2^{\pm 0.4}$	$22.8^{\pm 0.4}$	$21.1^{\pm 0.4}$	$19.6^{\pm 0.2}$	$18.6^{\pm 0.2}$
	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 %	$10.8^{\pm 0.1}$	$23.4^{\pm 0.4}$	$21.4^{\pm 0.1}$	$20.1^{\pm 0.3}$	$18.3^{\pm 0.5}$	$17.3^{\pm 0.2}$	$16.0^{\pm 0.1}$	$14.4^{\pm 0.1}$	$13.2^{\pm 0.2}$	$12.1^{\pm 0.1}$

Table 3.1: BLEU (avg±stddev) with transcription noise on ESIC dev set whose reference translations were English. The green-backgrounded area 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. In **bold** is near the maximum gap from single-source, more than 2.1 BLEU. Red-colored numbers are where at least one single-source scores higher.

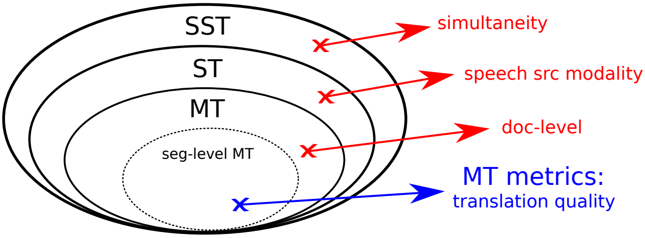


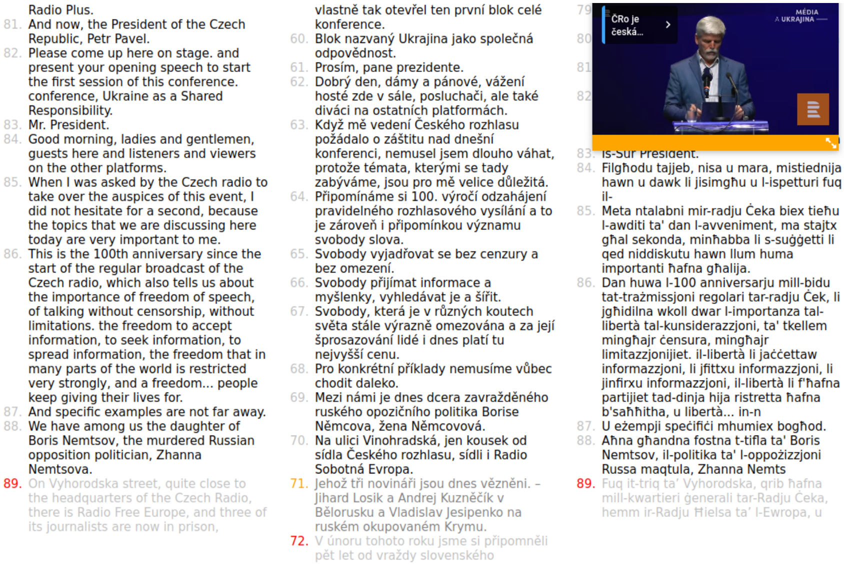
Figure 3.4: Illustration of simultaneous speech translation (SST) challenges that are not included in the design of machine translation (MT) metrics. We see simultaneous speech translation (SST) as a superset of offline speech translation (ST). The simultaneity lies in the difference. Then, we see ST as a superset of offline text-to-text machine translation (MT). Speech as source modality is in the difference. MT is a superset of the segment-level MT, which is designed to translate individual sentences, not documents that require document-level consistency that lies in the difference. The MT metrics are designed primarily for the segment-level MT, where the primary concern is the translation quality.

First, we needed to validate that the standard MT metrics that are designed for the offline text-to-text MT are reliable in SST, which has specific challenges illustrated in Figure 3.4. We compare human ratings of SST to the metrics and we realize that they strongly correlate. We used data from IWSLT 2022 shared task on English-to-German SST.

Second, we needed to know for our next experiments, what metric and reference to use: translation, or interpreting? We found out which metric setup correlates the most. The metric setup is the metric (COMET, BertScore, ChrF2, BLEU), reference (translation, interpreting, or both), and method for aligning the translation candidate to the reference, if the direct segment alignment is not available (mWERSegmenter, or concatenation to single sequence). We conclude with practical recommendations.

Whisper-Streaming For our further multi-source SST research in the realistic setup, we needed a state-of-the-art simultaneous ASR system. There was a robust and high-quality model Whisper (Radford et al., 2022), but it was available only for the offline mode. Independently of that, there was a state-of-the-art research of simultaneous streaming policies whose results and implementations remained mostly in research papers (Anastasopoulos et al., 2022; Polák et al., 2022). Therefore, we implemented a streaming mode for the Whisper model, which resulted in Whisper-Streaming, a very robust, high-quality and innovative demonstration tool of the simultaneous speech-to-text translation and transcription. We published the implementation[1] and it received lots of positive public

[1] https://github.com/ufal/whisper_streaming



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, 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.
87. And specific examples are not far away.
88. 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,

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 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.
66. Svobody přijímat informace a myšlenky, vyhledávat je a šířit.
67. 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, sídlí 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

79
80
81
82

83. Is-Sur President.
84. Filgħodu tajjeb, nisa u mara, mistiednija hawn u dawk li jisimgħu u l-ispetturi fuq il-
85. Meta ntalabni mir-radju Ċeka biex tieħu l-awditi ta' dan l-avveniment, ma stajtx għal sekonda, minħabba li s-suġġetti li qed niddiskutu hawn llum huma importanti ħafna għalija.
86. Dan huwa l-100 anniversarju mill-bidu tat-trażmissjoni regolari tar-radju Ċek, li jgħidilna wkoll dwar l-importanza tal-libertà tal-kunsiderazzjoni, ta' tkellem mingħajr ċensura, mingħajr limitazzjonijiet. il-libertà li jaċċettaw informazzjoni, li jfittxu informazzjoni, li jinfirxu informazzjoni, il-libertà li f'ħafna partijiet tad-dinja hija ristretta ħafna b'saħħitha, u libertà... in-n
87. U eżempji speċifiċi mhumiex bogħod.
88. Aħna għandna fostna t-tifla ta' Boris Nemtsov, il-politika ta' l-oppożizzjoni Russa maqtula, Zhanna Nemts
89. Fuq it-triq ta' Vyhorodska, qrib ħafna mill-kwartieri ġenerali tar-Radju Ċeka, hemm ir-Radju Ħielsa ta' l-Ewropa, u

Figure 3.5: A preview of presentation interface of simultaneous speech transcripts (middle column, in Czech) and translation (side columns, English and Maltese), in the ELITR presentation interface (Bojar et al., 2021b). The speech is a part of Czech Radio conference Media and Ukraine, 22nd June 2023.

interest. For example, GitHub documents 1 167 stars, 179 forks, and 10 contributors who delivered a new feature or a bug fix. We also integrated Whisper-Streaming into the ELITR live speech translation framework and demonstrated it in many real-life events, as illustrated on Figure 3.5.

4. Publications

In Table 4.1, we summarize our nine publications that are directly relevant to our thesis. We highlight four of them that contain our significant results.

Less related to our dissertation are another four publications in which we were involved during our PhD. program. There is a publication about the ELITR project (Bojar et al., 2020), and two publications about the ELITR complex and distributed system for live speech translation (Franceschini et al., 2020; Bojar et al., 2021a). We also co-authored the machine translation component in the CUNI system for IWSLT 2020 (Polák et al., 2020).

1.	A Speech Test Set of Practice Business Presentations with Additional Relevant Texts – Macháček, Kratochvíl, Vojtěchová and Bojar (2019) – Presented at SLSP 2019
2.	Presenting Simultaneous Translation in Limited Space – Macháček and Bojar (2020) – Presented at ITAT 2020
3.	ELITR Non-Native Speech Translation at IWSLT 2020 – Macháček, Kratochvíl, Sagar, Žilinec, Bojar, Nguyen, Schneider, Williams and Yao (2020) – Presented at IWSLT 2020
4.	**Lost in Interpreting: Speech Translation from Source or Interpreter?** – Macháček, Žilinec and Bojar (2021) – Presented at INTERSPEECH 2021
5.	The Reality of Multi-Lingual Machine Translation – Kocmi, Macháček and Bojar (2021) – book
6.	Continuous Rating as Reliable Human Evaluation of Simultaneous Speech Translation – Javorský, Macháček and Bojar (2022) – Presented at WMT 2022
7.	**Robustness of Multi-Source MT to Transcription Errors** – Macháček, Polák, Bojar and Dabre (2023c) – Published in Findings ACL 2023
8.	**MT Metrics Correlate with Human Ratings of Simultaneous Speech Translation** – Macháček, Bojar and Dabre (2023a) – Presented at IWSLT 2023
9.	**Turning Whisper into Real-Time Transcription System** – Macháček, Dabre and Bojar (2023b) – Presented at IJCNLP-AACL 2023 as system demonstration

Table 4.1: List of our publications relevant to this thesis or containing significant results reported in this thesis (highlighted in bold).

Acknowledgements

Funding for this work was provided by the following projects: H2020 ELITR (European Live Translator; Grant ID: 825460), GAUK Strojový překlad tlumočených projevů (Machine translation of interpreted speech; Grant ID: GAUK 398120), Babel Octopus: Robust Multi-Source Speech Translation (START Programme of Charles University; Grant ID: START/SCI/089), GA ČR Mnohojazyčný strojový překlad (Multi-Lingual Machine Translation; Grant ID: 18-24210S), GA ČR EXPRO NEUREM3 (Neural Representations in Multi-modal and Multi-lingual Modeling; Grant ID: 19-26934X, RIV: GX19-26934X), LINDAT/CLARIAH-CZ (Language Resources and Digital Arts and Humanities Research Infrastructure; Grant ID: LM2023062), and SVV 260 575 (Teoretické základy informatiky a výpočetní lingvistiky).

Bibliography

Anastasopoulos, A. et al. Findings of the IWSLT 2022 Evaluation Campaign. In Salesky, E. – Federico, M. – Costa-jussà, M. (Ed.) *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, p. 98–157, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.10. Available at: `https://aclanthology.org/2022.iwslt-1.10`.

Bojar, O. et al. ELITR: European Live Translator. In Martins, A. – Moniz, H. – Fumega, S. – Martins, B. – Batista, F. – Coheur, L. – Parra, C. – Trancoso, I. – Turchi, M. – Bisazza, A. – Moorkens, J. – Guerberof, A. – Nurminen, M. – Marg, L. – Forcada, M. L. (Ed.) *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, p. 463–464, Lisboa, Portugal, November 2020. European Association for Machine Translation. Available at: `https://aclanthology.org/2020.eamt-1.53`.

Bojar, O. et al. ELITR Multilingual Live Subtitling: Demo and Strategy. In Gkatzia, D. – Seddah, D. (Ed.) *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, p. 271–277, Online, April 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-demos.32. Available at: `https://aclanthology.org/2021.eacl-demos.32`.

Bojar, O. – Srdečný, V. – Kumar, R. – Smrž, O. – Schneider, F. – Haddow, B. – Williams, P. – Canton, C. Operating a Complex SLT System with Speakers and Human Interpreters. In Turchi, M. – Fantinuoli, C. (Ed.) *Proceedings of the 1st Workshop on Automatic Spoken Language Translation in Real-World Settings (ASLTRW)*, p. 23–34, Virtual, August 2021b. Association for Machine Translation in the Americas. Available at: `https://aclanthology.org/2021.mtsummit-asltrw.3`.

Dabre, R. – Cromieres, F. – Kurohashi, S. Enabling Multi-Source Neural Machine Translation By Concatenating Source Sentences In Multiple Languages. In Kurohashi, S. – Fung, P. (Ed.) *Proceedings of Machine Translation Summit XVI: Research Track*, p. 96–107, Nagoya Japan, September 18 – September 22 2017. Available at: `https://aclanthology.org/2017.mtsummit-papers.8`.

Firat, O. – Sankaran, B. – Al-onaizan, Y. – Yarman Vural, F. T. – Cho, K. Zero-Resource Translation with Multi-Lingual Neural Machine Translation. In Su, J. – Duh, K. – Carreras, X. (Ed.) *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 268–277, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1026. Available at: `https://aclanthology.org/D16-1026`.

Franceschini, D. et al. Removing European Language Barriers with Innovative Machine Translation Technology. In Rehm, G. – Bontcheva, K. – Choukri, K. – Hajič, J. – Piperidis, S. – Vasiļjevs, A. (Ed.) *Proceedings of the 1st International Workshop on Language Technology Platforms*, p. 44–49, Marseille, France, May 2020. European Language Resources Association. Available at: `https://aclanthology.org/2020.iwltp-1.7`. ISBN 979-10-95546-64-1.

Javorský, D. – Macháček, D. – Bojar, O. Continuous Rating as Reliable Human Evaluation of Simultaneous Speech Translation. In Koehn, P. et al. (Ed.) *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 154–164, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. Available at: `https://aclanthology.org/2022.wmt-1.9`.

Kocmi, T. – Macháček, D. – Bojar, O. *The Reality of Multi-Lingual Machine Translation.* ÚFAL, 2021.

Macháček, D. – Kratochvíl, J. – Vojtěchová, T. – Bojar, O. A Speech Test Set of Practice Business Presentations with Additional Relevant Texts. In Martín-Vide, C. – Purver, M. – Pollak, S. (Ed.) *Statistical Language and Speech Processing*, p. 151–161, Cham, 2019. Springer International Publishing. ISBN 978-3-030-31372-2.

Macháček, D. – Kratochvíl, J. – Sagar, S. – Žilinec, M. – Bojar, O. – Nguyen, T.-S. – Schneider, F. – Williams, P. – Yao, Y. ELITR Non-Native Speech Translation at IWSLT 2020. In Federico, M. – Waibel, A. – Knight, K. – Nakamura, S. – Ney, H. – Niehues, J. – Stüker, S. – Wu, D. – Mariani, J. – Yvon, F. (Ed.) *Proceedings of the 17th International Conference on Spoken Language Translation*, p. 200–208, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwslt-1.25. Available at: `https://aclanthology.org/2020.iwslt-1.25`.

Macháček, D. – Bojar, O. – Dabre, R. MT Metrics Correlate with Human Ratings of Simultaneous Speech Translation. In Salesky, E. – Federico, M. – Carpuat, M. (Ed.) *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, p. 169–179, Toronto, Canada (in-person and online), July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.iwslt-1.12. Available at: `https://aclanthology.org/2023.iwslt-1.12`.

Macháček, D. – Dabre, R. – Bojar, O. Turning Whisper into Real-Time Transcription System. In Saha, S. – Sujaini, H. (Ed.) *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, p. 17–24, Bali, Indonesia, November 2023b. Association for Computational Linguistics. Available at: `https://aclanthology.org/2023.ijcnlp-demo.3`.

Macháček, D. – Polák, P. – Bojar, O. – Dabre, R. Robustness of Multi-Source MT to Transcription Errors. In Rogers, A. – Boyd-Graber, J. – Okazaki, N. (Ed.) *Findings of the Association for Computational Linguistics: ACL 2023*, p. 3707–3723, Toronto, Canada, July 2023c. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.228. Available at: `https://aclanthology.org/2023.findings-acl.228`.

Macháček, D. – Bojar, O. Presenting Simultaneous Translation in Limited Space. In *Proceedings of the 20th Conference Information Technologies - Applications and Theory (ITAT 2020)*, p. 32–37, Košice, Slovakia, 2020. Tomáš Horváth.

Macháček, D. – Žilinec, M. – Bojar, O. Lost in Interpreting: Speech Translation from Source or Interpreter? In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association.* ISCA, 2021.

Martucci, G. – Cettolo, M. – Negri, M. – Turchi, M. Lexical Modeling of ASR Errors for Robust Speech Translation. In *Proc. Interspeech 2021*, p. 2282–2286, 2021. doi: 10.21437/Interspeech.2021-265.

Papi, S. – Gaido, M. – Negri, M. Direct Models for Simultaneous Translation and Automatic Subtitling: FBK@IWSLT2023. In Salesky, E. – Federico, M. – Carpuat, M. (Ed.) *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, p. 159–168, Toronto, Canada (in-person and online), July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.iwslt-1.11. Available at: `https://aclanthology.org/2023.iwslt-1.11`.

Papi, S. – Turchi, M. – Negri, M. AlignAtt: Using Attention-based Audio-Translation Alignments as a Guide for Simultaneous Speech Translation. In *Proc. INTERSPEECH 2023*, p. 3974–3978, 2023b. doi: 10.21437/Interspeech.2023-170.

Polák, P. – Sagar, S. – Macháček, D. – Bojar, O. CUNI Neural ASR with Phoneme-Level Intermediate Step for~Non-Native~SLT at IWSLT 2020. In Federico, M. – Waibel, A. – Knight, K. – Nakamura, S. – Ney, H. – Niehues, J. – Stüker, S. – Wu, D. – Mariani, J. – Yvon, F. (Ed.) *Proceedings of the 17th International Conference on Spoken Language Translation*, p. 191–199, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwslt-1.24. Available at: `https://aclanthology.org/2020.iwslt-1.24`.

Polák, P. – Pham, N.-Q. – Nguyen, T. N. – Liu, D. – Mullov, C. – Niehues, J. – Bojar, O. – Waibel, A. CUNI-KIT System for Simultaneous Speech Translation Task at IWSLT 2022. In Salesky, E. – Federico, M. – Costa-jussà, M. (Ed.) *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, p. 277–285, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.24. Available at: `https://aclanthology.org/2022.iwslt-1.24`.

Polák, P. – Liu, D. – Pham, N.-Q. – Niehues, J. – Waibel, A. – Bojar, O. Towards Efficient Simultaneous Speech Translation: CUNI-KIT System for Simultaneous Track at IWSLT 2023. In Salesky, E. – Federico, M. – Carpuat, M. (Ed.) *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, p. 389–396, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.iwslt-1.37. Available at: `https://aclanthology.org/2023.iwslt-1.37`.

Radford, A. – Kim, J. W. – Xu, T. – Brockman, G. – McLeavey, C. – Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision, 2022.

Ruiz, N. – Federico, M. Assessing the impact of speech recognition errors on machine translation quality. In Al-Onaizan, Y. – Simard, M. (Ed.) *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, p. 261–274, Vancouver, Canada, October 22-26 2014. Association for Machine Translation in the Americas. Available at: `https://aclanthology.org/2014.amta-researchers.20`.

Ruiz, N. – Gangi, M. A. D. – Bertoldi, N. – Federico, M. Assessing the Tolerance of Neural Machine Translation Systems Against Speech Recognition Errors. In *Proc. Interspeech 2017*, p. 2635–2639, 2017. doi: 10.21437/Interspeech.2017-1690.

Xue, H. – Feng, Y. – Gu, S. – Chen, W. Robust Neural Machine Translation with ASR Errors. In Wu, H. – Cherry, C. – Huang, L. – He, Z. – Liberman, M. – Cross, J. – Liu, Y. (Ed.) *Proceedings of the First Workshop on Automatic Simultaneous Translation*, p. 15–23, Seattle, Washington, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.autosimtrans-1.3. Available at: `https://aclanthology.org/2020.autosimtrans-1.3`.

Zoph, B. – Knight, K. Multi-Source Neural Translation. In Knight, K. – Nenkova, A. – Rambow, O. (Ed.) *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 30–34, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1004. Available at: `https://aclanthology.org/N16-1004`.