Long-form Simultaneous Speech Translation*
Thesis Proposal

Peter Polák
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
polak@ufal.mff.cuni.cz

Abstract

Simultaneous speech translation (SST) aims to provide real-time translation of spoken language, even before the speaker finishes their sentence. Traditionally, SST has been addressed primarily by cascaded systems that decompose the task into subtasks, including speech recognition, segmentation, and machine translation. However, the advent of deep learning has sparked significant interest in end-to-end (E2E) systems. Nevertheless, a major limitation of most approaches to E2E SST reported in the current literature is that they assume that the source speech is pre-segmented into sentences, which is a significant obstacle for practical, real-world applications. This thesis proposal addresses end-to-end simultaneous speech translation, particularly in the long-form setting, i.e., without pre-segmentation. We present a survey of the latest advancements in E2E SST, assess the primary obstacles in SST and its relevance to long-form scenarios, and suggest approaches to tackle these challenges.

1 Introduction

In today's highly globalized world, communication among individuals speaking different languages is gaining importance. International conferences and multinational organizations like the European Parliament often rely on human interpreters. However, in many scenarios, employing human interpreters can be impractical and costly. In such cases, simultaneous speech translation[1] (SST) offers a viable solution by enabling real-time translation before the speaker completes their sentence.

Traditionally, both offline speech translation (ST) and simultaneous speech translation (SST) have relied predominantly on cascaded systems that decompose the task into multiple subtasks, including speech recognition, speech segmentation, and machine translation (Osterholtz et al., 1992; Fügen et al., 2007; Bojar et al., 2021). However, recent advancements in deep learning and the availability of abundant data (Tan and Lim, 2018; Sperber and Paulik, 2020) have led to a significant paradigm shift towards end-to-end (E2E) models. While the cascaded approach continues to dominate offline ST, the opposite is true for SST (Anastasopoulos et al., 2022; Agarwal et al., 2023).

Despite the recent popularity of end-to-end SST, the vast majority of research focuses on the "short-form" setting, which assumes that the speech input is already pre-segmented into sentences. Critically, this assumption poses an obstacle to deployment in the wild. Therefore, we aim to achieve a "true" long-form simultaneous speech translation in our thesis. We break down our efforts into three steps:

Quality-latency tradeoff in SST The first step of our research concentrates on enhancing the quality-latency tradeoff, mainly in the traditional "short-form" regime. We will evaluate different approaches and architectures.

Towards the long-form SST In the next step, we will explore the feasibility of long-form simultaneous speech translation by adopting segmented inference.

True long-form SST The final goal of our work is to explore the potential of end-to-end modeling for true long-form SST. We will focus on identifying an appropriate model architecture and effective training procedures to achieve seamless and reliable long-form simultaneous speech translation.

The next section introduces some important aspects of simultaneous speech translation.

*The literature on simultaneous speech translation often uses the word "streaming" as an equivalent of "simultaneous" to refer to the translation of an unfinished utterance. In other literature, however, the term "streaming" refers to input spanning several sentences. To avoid confusion, we use "simultaneous" to refer to the translation of an unfinished utterance and "long-form" to refer to input spanning several sentences.

[1]We consider only the speech-to-text variant in this work.

2 Simultaneous Speech Translation

The ultimate goal of SST is to enable *real-time* communication between people speaking different languages. To achieve this goal, SST systems must meet two important criteria. First, they must be *computationally efficient* to ensure timely translation during ongoing speech. Second, SST systems must be capable of *handling unfinished sentences*. Working with unfinished sentences allows for more timely translations, particularly when waiting for sentences to be completed is impractical, such as matching slides or presenters' gestures. However, translating unfinished sentences increases the risk of translation errors since translation usually requires re-ordering that benefits from a more complete sentence context. Thus, there exists a *quality-latency tradeoff*. This means that given a certain latency constraint, we want the model to produce as good translations as possible. Ideally, we want the model to "predict" the future context without the risk of an incorrect translation. The quality-latency tradeoff is one of the main topics of our research.

2.1 Re-Translation vs. Incremental SST

SST can be classified as either re-translation or incremental. Re-translation SST (Niehues et al., 2016, 2018) can revise the hypothesis or re-rank the set of hypotheses as more speech input is read. Revising the translation allows the re-translation SST to have comparable final translation quality with the offline speech translation (Arivazhagan et al., 2020). This design approach arguably introduces challenges for the user in processing the translation and makes it impossible to use in real-time speech-to-speech translation. Additionally, it also complicates the latency evaluation.

In fact, several SST latency metrics (Ma et al., 2020) were originally developed specifically for incremental translation scenarios.[2] Incremental SST (Cho and Esipova, 2016; Dalvi et al., 2018) differs from the re-translation system in that it prunes all hypotheses to a common prefix, which is then shown to the user. For the user, the translation changes only by incrementally getting longer; none of the previously displayed outputs are ever modified. In our work, we focus on incremental SST.

[2]IWSLT shared tasks (Ansari et al., 2020; Anastasopoulos et al., 2021, 2022) also follow this evaluation standard.

2.2 Cascaded vs. End-to-End

Traditionally, offline speech translation and SST were achieved as a *cascade* of multiple systems: automatic speech recognition (ASR), inverse transcript normalization, which includes punctuation prediction and true casing, and machine translation (MT, Osterholtz et al., 1992; Fügen et al., 2007; Bojar et al., 2021). The advantage of the cascade approach is that we can optimize models for each subtask independently. Also, ASR and MT tasks typically have access to larger and more diverse corpora than direct speech translation.

However, using a cascade system introduces several challenges (Sperber and Paulik, 2020). The most important among them is *error propagation* (Ruiz and Federico, 2014). Further, MT models might suffer from *mismatched domains* when trained on written language. Furthermore, as the source is transformed into a textual form, it *loses crucial information about prosody*, i.e., the rhythm, intonation, and emphasis in speech (Bentivogli et al., 2021). Finally, many languages, especially endangered ones, have no written form, which makes the cascade approach impractical or impossible for such languages (Harrison, 2007; Duong et al., 2016).

As of the latest findings, the current state-of-the-art for offline speech translation continues to be based on a cascaded approach (Anastasopoulos et al., 2022; Agarwal et al., 2023). In simultaneous speech translation, however, both approaches yield competitive performance. The advantage of the end-to-end models in SST may be that they avoid the extra delay caused by ASR-MT collaboration in the cascade (Wang et al., 2022).

In our work, we focus on end-to-end models.

3 Long-form Simultaneous Speech Translation

Most of the contemporary research on SST assumes speech pre-segmented into short utterances with segmentation following the sentence boundaries. However, in any real application, there is no such segmentation available. This section places long-form SST within the broader context of long-form ASR, MT, and offline ST. Subsequently, we explore the current literature on long-form SST.

3.1 Long-Form ASR

In terms of input and output modalities, long-form ASR and ST face similar issues. There are two

types of strategies for long-form processing: (1) the *segmented approach*, which divides the input into smaller chunks, and (2) the *true long-form approach*, which handles the entire long-form input as a single unit.

Most of the literature focuses on the *segmented approach*. A typical solution involves pre-segmenting the audio using voice activity detection (VAD). However, VAD segmentation may not be optimal for real-world speech since it might fail to handle hesitations or pauses in sentences that must be treated as undivided units. More sophisticated approaches leverage latent alignments obtained from CTC (Graves et al., 2006) and RNN-T (Graves, 2012) for better segmentation (Yoshimura et al., 2020; Huang et al., 2022). Alternatively, segmentation into *fixed segments* is also popular (Chiu et al., 2019, 2021). To reduce low-quality transcripts close to the segment boundaries, they typically perform overlapped inference and use latent alignments to merge the transcripts correctly. The chunking approach is also adopted by the attentional model Whisper in the offline (Radford et al., 2023) and simultaneous regime (Macháček et al., 2023).

Another line of work focused on *long-form modeling* directly. For example, Chiu et al. (2019) conducted a comprehensive study comparing different architectures, including RNN-T and attention-based models. The findings indicate that only RNN-T and CTC architectures can generalize to unseen lengths. To further improve the true long-form ASR, Narayanan et al. (2019) suggest simulation of long-form training by LSTM state passing.

While the previously mentioned research was predominantly based on RNNs, more recent work has transitioned to utilizing Transformer models. Zhang et al. (2023) compared a chunk-wise attention encoder, which involves an encoder with a limited attention span, in combination with the attention-based decoder (AD) and CTC. We note that while the encoder has a limited attention span, the attention-based decoder sees the entire encoder representation. The model employing AD could not function without chunking, whereas the CTC model processed the entire speech at once and still outperformed the AD model.

3.2 Long-Form MT

The primary objective of long-form MT is to enhance textual coherence, as conventional MT systems assume sentence independence. Early work explored a concatenation of previous (Tiedemann and Scherrer, 2017; Donato et al., 2021) and future sentences (Agrawal et al., 2018). These works showed that MT models benefit from the extra context and better handle the inter-sentential discourse phenomena. However, the benefits diminish if the context grows beyond a few sentences (Agrawal et al., 2018; Kim et al., 2019; Fernandes et al., 2021). This can be attributed to the limitations of attention mechanisms, where an extensive volume of irrelevant information can lead to confusion.

Other body of work tries to model very long sequences directly. Dai et al. (2019) introduced a recurrence mechanism and improved positional encoding scheme in the Transformer. Later work proposed an explicit compressed memory realized by a few dense vectors (Feng et al., 2022).

3.3 Long-Form Offline ST

Unlike written input text in long-form MT, speech input in the ST task lacks explicit information about segmentation. Therefore, the research in the area of long-form offline speech translation concentrates on two separate issues: (1) improving *segmentation* into sentences, and (2) enhancing robustness through the use of larger *context*.

In the traditional cascaded approach with separate speech recognition and machine translation models, the work focused on segmentation strategies for the ASR transcripts.[3] The methods are usually based on re-introducing punctuation to the transcript (Lu and Ng, 2010; Rangarajan Sridhar et al., 2013; Cho et al., 2015, 2017). However, these approaches suffer from ASR error propagation and disregard the source audio's acoustic information. This was addressed by Iranzo-Sánchez et al. (2020a), however, the approach still requires an intermediate ASR transcript that is unavailable in E2E models.

An alternative approach involves source-speech-based segmentation. The early work focused on VAD segmentation. This is usually sub-optimal as speakers place pauses inside sentences, not necessarily between them (e.g., hesitations before words with high information content, Goldman-Eisler, 1958). To this end, researchers tried considering not only the presence of speech but also its length (Potapczyk and Przybysz, 2020; Inaguma et al.,

[3]ASR transcripts are traditionally normalized, i.e., they consist of lowercase words without punctuation.

2021; Gaido et al., 2021). Later studies tried to avoid VAD and focused on more linguistically-motivated approaches, e.g., ASR CTC to predict voiced regions Gállego et al. (2021) or directly modeling the sentence segmentation (Tsiamas et al., 2022b; Fukuda et al., 2022).

To address the problem of inadequate segmentation, Gaido et al. (2020) showed that context-aware ST is less prone to segmentation errors. In an extensive study of context-aware ST, Zhang et al. (2021) observed that context improves quality, but this holds only for a limited number of utterances.

3.4 Long-Form Simultaneous ST

Research focusing on direct long-form simultaneous speech translation remains relatively scarce. The closest works are in long-form simultaneous MT. Schneider and Waibel (2020) proposed a streaming MT model capable of translating unsegmented text input. This model could be theoretically adapted for speech input. However, it was later shown that this model exhibits huge latency (Iranzo Sanchez et al., 2022). Another work (Iranzo Sanchez et al., 2022) explored the extended context and confirmed the findings from long-form MT and offline ST, demonstrating that using the previous context significantly enhances performance. They also confirmed that a too-long context leads to decreased translation quality.

Finally, the only direct SST model that claims to work on a possibly unbounded input is Ma et al. (2021). The model utilizes a Transformer encoder with a restriction on self-attention, allowing it to attend solely to a memory bank and a small segment. Unfortunately, based on the reported experiments, whether the model was specifically evaluated in the long-form setting remains unclear.

3.5 Evaluation

Evaluation of SST is a complex problem as we have to consider not only the translation quality but also the latency. Additionally, in the long-form regime, segmentation becomes another obstacle.

The most commonly used metric for translation quality in speech translation is BLEU (Papineni et al., 2002; Post, 2018). Other metrics such as chrF++ (Popović, 2017) and a neural-based metric COMET (Rei et al., 2020) can be applied, too.

The other important property of an SST system is latency. There are two main types of latencies: computation-unaware (CU) and computation-aware (CA) latency. The computation-unaware latency measures the delay in emitting a translation token relative to the source, regardless of the actual computation time. Hence, CU latency allows for a fair comparison regardless of the hardware infrastructure. However, CU latency cannot penalize the evaluated system for extensive computation; hence, CA latency can offer a more realistic assessment.

Measuring latency relative to the source or reference in SST is quite difficult because of the reordering present in translation. Historically, latency metrics were first developed for simultaneous machine translation (i.e., the source is text rather than speech). The most common are average lagging (AL; Ma et al., 2019) and differentiable average lagging (DAL; Cherry and Foster, 2019). Broadly speaking, they measure "how much of the source was read by the system to translate a word". The latency unit is typically a word. The speech community quickly adopted these metrics. Unfortunately, these metrics assume a uniform distribution of words and uniform length of these words in the speech source. Alternatively, Ansari et al. (2021) proposed to use a statistical word alignment of the candidate translation with the corresponding source transcript. This theoretically allows for more precise latency evaluation, but it is unclear how the alignment errors impact the reliability.

In the unsegmented long-form setting, additional issues arise. In a typical "short-form" segmented setup, the SST model does inference on a pre-segmented input. However, the candidate and reference segmentation into sentences might differ in the long-form unsegmented regime. Traditionally, this issue was addressed by re-segmenting the hypothesis based on the reference (Matusov et al., 2005). After the re-segmentation, a standard sentence-level evaluation of translation quality and latency is done. It should be noted that the commonly used latency metrics (AL, DAL) cannot be used in the long-form regime (Iranzo-Sánchez et al., 2021) without the re-segmentation. Yet, recent work observed that the re-segmentation introduces errors (Amrhein and Haddow, 2022). This poses a risk of incorrect translation and quality assessment and remains an open research question.

4 Thesis Goals

The goal of our thesis is to achieve a "true" long-form simultaneous speech translation. This section outlines the steps we will take to accomplish this goal.

4.1 Data and Evaluation

In our future research, we will mainly use the setup similar to the IWSLT shared tasks (Ansari et al., 2020; Anastasopoulos et al., 2021, 2022), i.e., mostly single speaker data. Identical to the IWSLT, we will treat the TED data as an in-domain setting. We will consider domains such as parliamentary speeches (e.g., Europarl-ST Iranzo-Sánchez et al., 2020b) for the out-of-domain setting. As for the languages, we will include a diverse set of language pairs. A good inspiration might be again the IWSLT, i.e., English-to-{German, Japanese, Chinese}. Challenging will be the long-form setting, as to the best of our knowledge, none of the available data is strictly long-form. Our preliminary review found that the original TED talks can be reconstructed from the MuST-C (Cattoni et al., 2021) development and test set available for English-to-{German, Japanese, Chinese} language pairs.

As highlighted in the literature review in Section 3.5, evaluating the long-form SST remains an open problem. The quality and latency evaluation metrics currently used are designed for sentence-level evaluation. We must re-segment the long hypotheses into sentences based on their word alignment with provided references to use these metrics in the long-form regime. Unfortunately, the re-segmentation introduces errors, which poses a risk to the evaluation reliability. To tackle this, we will investigate alternative evaluation strategies. One potential approach for reducing the alignment error could be to move the alignment to the sentence level rather than the word level and allow an m-to-n mapping between the reference and proposed sentences, similar to the Gale–Church alignment algorithm (Gale et al., 1994), with a reasonably small m and n (e.g., $0 \leq m, n \leq 2$). To verify the effectiveness of this method, we need to compare its correlation with human evaluations.

4.2 Quality-latency tradeoff in SST

The first step of our research concentrates on enhancing the quality-latency tradeoff, mainly in the traditional "short-form" simultaneous speech translation. We hope the insights and improvements from the short-form regime will translate into the long-form regime.

In the research done so far, we already successfully reviewed the possibility of "onlinizing" state-of-the-art offline speech translation models in Polák et al. (2022). Our observations indicated that the attention-based encoder-decoder (AED) models tend to over-generate. This not only affects the resulting quality but also negatively impacts the AL latency evaluation reliability. Therefore, we proposed an improved version of the AL metric, which was later independently proposed under name length-adaptive average lagging (LAAL; Papi et al., 2022). To remedy the over-generation problem, we proposed an improved version of the beam search algorithm in Polák et al. (2023b). While this led to significant improvements in the quality-latency tradeoff, the decoding still relied on label-synchronous decoding. In Polák et al. (2023a), we proposed a novel SST policy dubbed "CTC policy" that uses the output of an auxiliary CTC layer to guide the decoding. The proposed CTC policy led to even greater improvements in quality and reduced the real-time factor to 50 %.

Thus far, our research has focused primarily on the AED architecture. Nonetheless, recent findings (Anastasopoulos et al., 2022; Agarwal et al., 2023) suggest that other approaches, such as transducers (Graves, 2012), yield competitive results. Nevertheless, it remains unclear which approach is the most advantageous for SST. Our goal will be to compare these architectures for SST. We will put a particular emphasis on architectures with latent alignments (e.g., transducers). Generally, the latent alignment models make a strong monotonic assumption on the mapping between the source and the target, which might be problematic for the translation, typically involving word reordering. Therefore, we will assess the alignment quality and potential applications (such as segmentation).

4.3 Towards the Long-Form SST via On-the-Fly Segmentation

In the second stage, we will concentrate on the long-form SST by utilizing on-the-fly segmentation and short-form models from the previous stage.

Drawing inspiration from offline long-form ST, which primarily emphasizes segmentation, we consider direct segmentation modeling the most promising approach (Tsiamas et al., 2022a; Fukuda et al., 2022). The limitation of these approaches is that they do not allow out-of-the-box simultaneous inference. However, we believe their adaptation to the simultaneous regime should be relatively straightforward (e.g., using a unidirectional encoder) and a custom decoding strategy. The main challenge here will be integrating this segmenta-

tion with existing models, especially considering the quality-latency tradeoff.

Our hopes go even further: Can we train a model to translate and predict the segmentation at the same time? The translation already contains punctuation marks (full stop, exclamation, and question marks), so if we knew the alignment between the translation and the source speech, we could use this information to segment the utterances directly. Therefore, we will experiment with various alignment approaches and asses their applicability to the segmentation. The results of our initial investigation on on-the-fly separation with CTC outputs are available in Polák and Bojar (2023).

However, we see another valuable use of direct speech-to-translation alignments — dataset creation. Today, ST datasets are created using the cascaded approach (Iranzo-Sánchez et al., 2020b; Cattoni et al., 2021; Salesky et al., 2021). The source transcript is first forced-aligned to the speech, then the transcript is word-aligned to the translations, and finally, these two alignments are used to segment the source speech into sentences based on the punctuation in the translation. In fact, this approach has a critical drawback: it virtually eliminates all data without a source transcript, preventing the research community from utilizing potentially valuable data sources. It is also worth noting that some languages do not have a writing system, which makes the direct speech-to-translation alignment even more attractive. Therefore, if the alignments show promising results, we will explore the feasibility of E2E speech-to-translation dataset creation.

An additional question is how to accommodate long context in the simultaneous regime. As pointed out in Sections 3.2 to 3.4, the performance usually drops with a context longer than a few sentences. Some solutions have been suggested (Kim et al., 2019; Feng et al., 2022), but it remains unclear how to adapt these approaches for SST with the specifics of SST in mind (e.g., computational constraints, speech input).

4.4 True Long-Form SST

The ultimate goal of our work is to achieve true long-form simultaneous speech translation. In other words, we aim to develop an architecture capable of processing a potentially infinite stream of speech input without any segmentation or special inference algorithm, translating the speech directly into the target language in real time. Admittedly, this is a very ambitious goal. However, there is plenty of evidence that it is feasible. For example, in long-form ASR, related work has already observed that the RNN-T and CTC architectures are capable of long-form regime (Chiu et al., 2019; Narayanan et al., 2019; Lu et al., 2021; Zhang et al., 2023; Rekesh et al., 2023). Arguably, speech recognition is simpler than speech translation because it monotonically transcribes speech without reordering. However, the literature also shows that an architecture like RNN-T can be used in the "short-form" offline and simultaneous ST (Yan et al., 2023).

Therefore, based on the previous work in speech recognition and translation, we will propose a novel architecture that will allow simultaneous speech translation of a possibly infinite stream of speech. We will take inspiration from the existing architectures but revise them for the specific needs of simultaneous ST. This will require a particular focus on speech-to-translation alignment so that the source speech and target translation do not get out of sync. This architecture will also contain a "forgetting" mechanism that will allow the storage of essential bits of context while preventing memory issues. Finally, we will address the train-test mismatch because current hardware and training methods do not permit models to fit long inputs.

5 Conclusion

In conclusion, this thesis proposal presents an overview of the challenges involved in simultaneous speech translation (SST). The literature review highlighted the limited research on long-form speech translation. Our research sets out three main goals with an emphasis on long-form speech translation. These include improving the general quality-latency tradeoff in SST, exploring long-form SST through segmented inference, and ultimately achieving true long-form SST modeling. We placed these goals in the context of related work and outlined a clear strategy for achieving them.

Acknowledgments

Peter would like to thank his supervisor, Ondřej Bojar, for his insight and guidance, as well as the anonymous reviewers for their valuable suggestions. This work has received support from GAUK project 244523 of Charles University and partial support from grant 19-26934X (NEUREM3) of the Czech Science Foundation.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Ruchit Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 31–40, Alicante, Spain.

Chantal Amrhein and Barry Haddow. 2022. Don't discard fixed-window audio segmentation in speech-to-text translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 203–219, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. 2021. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online). Association for Computational Linguistics.

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. 2020. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online. Association for Computational Linguistics.

Ebrahim Ansari, Ondřej Bojar, Barry Haddow, and Mohammad Mahmoudi. 2021. SLTEV: Comprehensive evaluation of spoken language translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 71–79, Online. Association for Computational Linguistics.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. 2020. Re-translation versus streaming for simultaneous translation. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 220–227, Online. Association for Computational Linguistics.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2873–2887, Online. Association for Computational Linguistics.

Ondřej Bojar, Dominik Macháček, Sangeet Sagar, Otakar Smrž, Jonáš Kratochvíl, Peter Polák, Ebrahim Ansari, Mohammad Mahmoudi, Rishu Kumar, Dario Franceschini, Chiara Canton, Ivan Simonini, Thai-Son Nguyen, Felix Schneider, Sebastian Stüker, Alex Waibel, Barry Haddow, Rico Sennrich, and Philip Williams. 2021. ELITR multilingual live subtitling: Demo and strategy. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 271–277, Online. Association for Computational Linguistics.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Must-c: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155.

Colin Cherry and George Foster. 2019. Thinking slow about latency evaluation for simultaneous machine translation. *arXiv preprint arXiv:1906.00048*.

Chung-Cheng Chiu, Wei Han, Yu Zhang, Ruoming Pang, Sergey Kishchenko, Patrick Nguyen, Arun Narayanan, Hank Liao, Shuyuan Zhang, Anjuli Kannan, et al. 2019. A comparison of end-to-end models for long-form speech recognition. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 889–896. IEEE.

Chung-Cheng Chiu, Arun Narayanan, Wei Han, Rohit Prabhavalkar, Yu Zhang, Navdeep Jaitly, Ruoming Pang, Tara N Sainath, Patrick Nguyen, Liangliang Cao, et al. 2021. Rnn-t models fail to generalize to out-of-domain audio: Causes and solutions. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 873–880. IEEE.

Eunah Cho, Jan Niehues, Kevin Kilgour, and Alex Waibel. 2015. Punctuation insertion for real-time spoken language translation. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Papers*, pages 173–179.

Eunah Cho, Jan Niehues, and Alex Waibel. 2017. Nmt-based segmentation and punctuation insertion for real-time spoken language translation. In *Interspeech*, pages 2645–2649.

Kyunghyun Cho and Masha Esipova. 2016. Can neural machine translation do simultaneous translation? *arXiv preprint arXiv:1606.02012*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, and Stephan Vogel. 2018. Incremental decoding and training methods for simultaneous translation in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 493–499, New Orleans, Louisiana. Association for Computational Linguistics.

Domenic Donato, Lei Yu, and Chris Dyer. 2021. Diverse pretrained context encodings improve document translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1299–1311, Online. Association for Computational Linguistics.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 949–959, San Diego, California. Association for Computational Linguistics.

Yukun Feng, Feng Li, Ziang Song, Boyuan Zheng, and Philipp Koehn. 2022. Learn to remember: Transformer with recurrent memory for document-level machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1409–1420, Seattle, United States. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics.

Christian Fügen, Alex Waibel, and Muntsin Kolss. 2007. Simultaneous translation of lectures and speeches. *Machine translation*, 21:209–252.

Ryo Fukuda, Katsuhito Sudoh, and Satoshi Nakamura. 2022. Speech segmentation optimization using segmented bilingual speech corpus for end-to-end speech translation. *arXiv preprint arXiv:2203.15479*.

Marco Gaido, Mattia A. Di Gangi, Matteo Negri, Mauro Cettolo, and Marco Turchi. 2020. Contextualized Translation of Automatically Segmented Speech. In *Proc. Interspeech 2020*, pages 1471–1475.

Marco Gaido, Matteo Negri, Mauro Cettolo, and Marco Turchi. 2021. Beyond voice activity detection: Hybrid audio segmentation for direct speech translation. In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 55–62.

William A. Gale, Kenneth Ward Church, et al. 1994. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102.

Gerard I. Gállego, Ioannis Tsiamas, Carlos Escolano, José A. R. Fonollosa, and Marta R. Costa-jussà. 2021. End-to-end speech translation with pre-trained models and adapters: UPC at IWSLT 2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 110–119, Bangkok, Thailand (online). Association for Computational Linguistics.

Frieda Goldman-Eisler. 1958. Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology*, 10(2):96–106.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

K David Harrison. 2007. *When languages die: The extinction of the world's languages and the erosion of human knowledge*. Oxford University Press.

W Ronny Huang, Shuo-yiin Chang, David Rybach, Rohit Prabhavalkar, Tara N Sainath, Cyril Allauzen, Cal Peyser, and Zhiyun Lu. 2022. E2e segmenter: Joint segmenting and decoding for long-form asr. *arXiv preprint arXiv:2204.10749*.

Hirofumi Inaguma, Brian Yan, Siddharth Dalmia, Pengcheng Guo, Jiatong Shi, Kevin Duh, and Shinji Watanabe. 2021. ESPnet-ST IWSLT 2021 offline speech translation system. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 100–109, Bangkok, Thailand (online). Association for Computational Linguistics.

Javier Iranzo Sanchez, Jorge Civera, and Alfons Juan-Císcar. 2022. From simultaneous to streaming machine translation by leveraging streaming history. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6972–6985, Dublin, Ireland. Association for Computational Linguistics.

Javier Iranzo-Sánchez, Jorge Civera Saiz, and Alfons Juan. 2021. Stream-level latency evaluation for simultaneous machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 664–670, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Javier Iranzo-Sánchez, Adrià Giménez Pastor, Joan Albert Silvestre-Cerdà, Pau Baquero-Arnal, Jorge Civera Saiz, and Alfons Juan. 2020a. Direct segmentation models for streaming speech translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2599–2611, Online. Association for Computational Linguistics.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerda, Javier Jorge, Nahuel Roselló, Adria Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020b. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233. IEEE.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, page 24–34, Hong Kong, China. Association for Computational Linguistics.

Wei Lu and Hwee Tou Ng. 2010. Better punctuation prediction with dynamic conditional random fields. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 177–186.

Zhiyun Lu, Yanwei Pan, Thibault Doutre, Parisa Haghani, Liangliang Cao, Rohit Prabhavalkar, Chao Zhang, and Trevor Strohman. 2021. Input length matters: Improving rnn-t and mwer training for long-form telephony speech recognition. *arXiv preprint arXiv:2110.03841*.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Xutai Ma, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. 2020. SIMULEVAL: An evaluation toolkit for simultaneous translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 144–150, Online. Association for Computational Linguistics.

Xutai Ma, Yongqiang Wang, Mohammad Javad Dousti, Philipp Koehn, and Juan Pino. 2021. Streaming simultaneous speech translation with augmented memory transformer. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7523–7527. IEEE.

Dominik Macháček, Raj Dabre, and Ondřej Bojar. 2023. Turning whisper into real-time transcription system. In *Proceedings of the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 13th International Joint Conference on Natural Language Processing: System Demonstrations*, Bali, Indonesia. Association for Computational Linguistics.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

Arun Narayanan, Rohit Prabhavalkar, Chung-Cheng Chiu, David Rybach, Tara N Sainath, and Trevor Strohman. 2019. Recognizing long-form speech using streaming end-to-end models. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 920–927. IEEE.

J. Niehues, T. S. Nguyen, E. Cho, T.-L. Ha, K. Kilgour, M. Müller, M. Sperber, S. Stüker, and A. Waibel. 2016. Dynamic transcription for low-latency speech

translation. In *17th Annual Conference of the International Speech Communication Association, INTERSPEECH 2016; Hyatt Regency San FranciscoSan Francisco; United States; 8 September 2016 through 16 September 2016*, volume 08-12-September-2016 of *Proceedings of the Annual Conference of the International Speech Communication Association. Ed. : N. Morgan*, pages 2513–2517. International Speech Communication Association.

J. Niehues, N.-Q. Pham, T.-L. Ha, M. Sperber, and A. Waibel. 2018. Low-latency neural speech translation. In *19th Annual Conference of the International Speech Communication, INTERSPEECH 2018; Hyderabad International Convention Centre (HICC)Hyderabad; India; 2 September 2018 through 6 September 2018. Ed.: C.C. Sekhar*, volume 2018-September of *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1293–1297. ISCA.

L. Osterholtz, C. Augustine, A. McNair, I. Rogina, H. Saito, T. Sloboda, J. Tebelskis, and A. Waibel. 1992. Testing generality in janus: a multi-lingual speech translation system. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 209–212 vol.1.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2022. Over-generation cannot be rewarded: Length-adaptive average lagging for simultaneous speech translation. In *Proceedings of the Third Workshop on Automatic Simultaneous Translation*, pages 12–17, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Peter Polák and Ondřej Bojar. 2023. Long-form end-to-end speech translation via latent alignment segmentation. *arXiv preprint arXiv:2309.11384*.

Peter Polák, Danni Liu, Ngoc-Quan Pham, Jan Niehues, Alexander Waibel, and Ondřej Bojar. 2023a. Towards efficient simultaneous speech translation: CUNI-KIT system for simultaneous track at IWSLT 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 389–396, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Peter Polák, Ngoc-Quan Pham, Tuan Nam Nguyen, Danni Liu, Carlos Mullov, Jan Niehues, Ondřej Bojar, and Alexander Waibel. 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.

Peter Polák, Brian Yan, Shinji Watanabe, Alexander Waibel, and Ondrej Bojar. 2023b. Incremental Blockwise Beam Search for Simultaneous Speech Translation with Controllable Quality-Latency Tradeoff. In *Proc. Interspeech 2023*.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Tomasz Potapczyk and Pawel Przybysz. 2020. SR-POL's system for the IWSLT 2020 end-to-end speech translation task. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 89–94, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore, Andrej Ljolje, and Rathinavelu Chengalvarayan. 2013. Segmentation strategies for streaming speech translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 230–238, Atlanta, Georgia. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Dima Rekesh, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Juang, Oleksii Hrinchuk, Ankur Kumar, and Boris Ginsburg. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. *arXiv preprint arXiv:2305.05084*.

Nicholas Ruiz and Marcello Federico. 2014. Assessing the impact of speech recognition errors on machine translation quality. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, pages 261–274.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W Oard, and Matt Post. 2021. The multilingual tedx corpus for speech recognition and translation. *arXiv preprint arXiv:2102.01757*.

Felix Schneider and Alexander Waibel. 2020. Towards stream translation: Adaptive computation time for simultaneous machine translation. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 228–236, Online. Association for Computational Linguistics.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, Online. Association for Computational Linguistics.

Kar-Han Tan and Boon Pang Lim. 2018. The artificial intelligence renaissance: deep learning and the road to human-level machine intelligence. *APSIPA Transactions on Signal and Information Processing*, 7:e6.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Ioannis Tsiamas, Gerard I. Gállego, Carlos Escolano, José Fonollosa, and Marta R. Costa-jussà. 2022a. Pretrained speech encoders and efficient finetuning methods for speech translation: UPC at IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 265–276, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Ioannis Tsiamas, Gerard I Gállego, José AR Fonollosa, and Marta R Costa-jussà. 2022b. Shas: Approaching optimal segmentation for end-to-end speech translation. *arXiv preprint arXiv:2202.04774*.

Minghan Wang, Jiaxin Guo, Yinglu Li, Xiaosong Qiao, Yuxia Wang, Zongyao Li, Chang Su, Yimeng Chen, Min Zhang, Shimin Tao, Hao Yang, and Ying Qin. 2022. The HW-TSC's simultaneous speech translation system for IWSLT 2022 evaluation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 247–254, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Brian Yan, Jiatong Shi, Yun Tang, Hirofumi Inaguma, Yifan Peng, Siddharth Dalmia, Peter Polak, Patrick Fernandes, Dan Berrebbi, Tomoki Hayashi, Xiaohui Zhang, Zhaoheng Ni, Moto Hira, Soumi Maiti, Juan Pino, and Shinji Watanabe. 2023. ESPnet-ST-v2: Multipurpose spoken language translation toolkit. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–411, Toronto, Canada. Association for Computational Linguistics.

Takenori Yoshimura, Tomoki Hayashi, Kazuya Takeda, and Shinji Watanabe. 2020. End-to-end automatic speech recognition integrated with ctc-based voice activity detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6999–7003. IEEE.

Biao Zhang, Ivan Titov, Barry Haddow, and Rico Sennrich. 2021. Beyond sentence-level end-to-end speech translation: Context helps. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2566–2578.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.