Simultaneous and Long-form 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

The goal of the simultaneous speech translation (SST) is to provide real-time translation before the speaker completes the sentence. Traditionally, SST has been addressed primarily by cascaded systems that decompose the task into subtasks, including speech recognition, speech 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 focuses on end-to-end simultaneous speech translation, especially in the long-form setting, i.e., not pre-segmented. We provide an overview of recent developments in the field of E2E SST, examine the main challenges associated with SST and its applicability in long-form scenarios, and propose methods to address these challenges.

1 Introduction

In today's highly globalized world, communication among individuals speaking different languages is gaining importance. International conferences and multinational organizations, such as 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 (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; Müller et al., 2016; 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 lead to a significant paradigm shift towards end-to-end (E2E) models. While the cascaded approach continues to dominate in 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 within the research community, 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.

In this thesis proposal, we present the goal of our research — simultaneous and long-form speech translation. Our plan starts with the above-mentioned "short-form" assumption, which allows us to revisit some modeling and algorithmic approaches. We then gradually introduce the long-form regime into our experiments by focusing on leveraging models from the "short-form" SST and on-the-fly segmentation. Finally, our ultimate goal is a direct end-to-end long-form speech translation.

This thesis proposal is structured as follows: we begin with a gentle description of simultaneous speech translation in Section 2. This is followed by an overview of the most important architectures for speech processing in Section 3. In Section 4, we introduce the long-form regime. We highlight the difficulties and review the literature on related long-form tasks. In Section 5, we outline the goals of our research. In Section 6, we present our previous work, and in Section 7, we present our future plans.

* 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" is used to refer to input that spans several sentences. To avoid confusion, we have chosen the word "simultaneous" to refer to the translation of an unfinished utterance, and "long-form" to refer to input that spans several sentences.

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 to meet two important criteria. First, they must be *computationally efficient* to ensure timely translation during ongoing speech. To address this intricate problem, the deep learning community has proposed various techniques, including model pruning (Reed, 1993), model quantization (Gong et al., 2014), and knowledge distillation (Hinton et al., 2015). Second, SST systems must be *capable of handling unfinished sentences*. Working with unfinished sentences allows for more timely translations, particularly in scenarios where 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 reordering 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, but without the risk of an incorrect translation.

The *quality-latency tradeoff* in SST is one of the main topics of our research. We review the key aspects of this topic in the following section.

2.1 Human Translation and Interpreting vs. SST

Human translation and interpreting represent two different approaches to facilitating communication between speakers of different languages. Understanding these two tasks will allow us to better understand the SST task, its commonalities, and key differences compared to human translation and interpreting.

Translation refers to the task of reformulating of a written source text into a written target language text, while *interpreting* refers to the non-written re-expression of a non-written source (Gile, 2004). In the context of this work, we understand the non-written source as speech.

Translation and interpreting differ in several key aspects (Gile, 2004). First, interpreters face strict time constraints. They often have only seconds to process the information and deliver the translation, whereas the translators typically have more time for the translation. Another key aspect is the target medium. Translations are presented as text that can be reviewed and read at any pace. In contrast, interpreting delivers its end product in an oral form. The interpretation must follow the pace of the source, but also be understood by the target audience. These considerations are reflected in the contrasting goals of translation and interpreting: while the translation favors accuracy, interpreting focuses on conveying the core meaning within a limited time. Interpreters must make quick decisions about word choice, phrasing, and contextual adaptation to facilitate seamless communication in real time. While both translators and interpreters may provide explanations of the intercultural context, interpreters may also comment on the actions of other actors (e.g., on stage), provide organizational comments, or interpret other relevant sources such as slides or documents. Finally, because interpreters must to listen, process the information, and deliver the interpretation in real time, they are under a high cognitive load, which in turn can lead to exhaustion. In some cases, exhaustion can force interpreters to omit parts of the source, i.e., making the interpretation less reliable.

In the context of human translation and interpreting, simultaneous speech translation borrows from both disciplines. Similar to the translation, SST typically produces written text in the target language. In addition, SST tries to provide accurate translations without considering the complexity of the text (e.g., avoids simplifications). On the other hand, similar to interpreting, SST transforms the source into speech. SST also tries to produce the translations in real time. However, unlike interpreting, SST cannot be exhausted and therefore can therefore be considered more reliable in terms of source coverage.

While contemporary SST systems have not yet reached the quality of human interpreters in terms of handling out-of-domain topics[1] and dealing with multimodality, they prove invaluable in situations where employing human interpreters is impractical or cost prohibitive.

2.2 Re-Translation vs. Incremental SST

SST can be classified as either re-translation or incremental (see Figure 1). Re-translation SST (Niehues et al., 2016, 2018) maintains multiple hy-

[1] See human quality evaluation of out-of-domain non-native speeches translated/interpreted by SST and a human interpreter in Anastasopoulos et al. (2022); Agarwal et al. (2023).

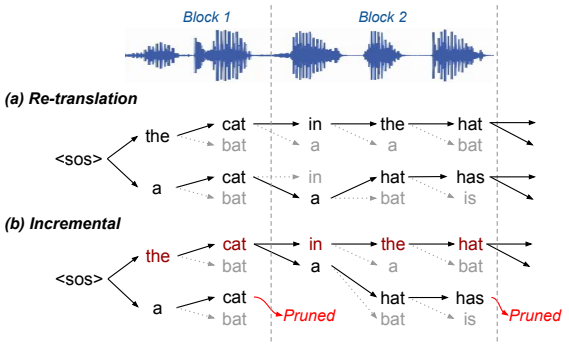

Figure 1: Blockwise re-translation vs. incremental SST. Decoding is advanced after a new block of speech becomes available (here: the dashed lines). The *re-translation SST can revise its previous translation* presented to the user, allowing it to maintain multiple hypotheses with different prefixes. In contrast, the *incremental SST cannot revise its translation*, and is limited to maintaining only one hypothesis after each incremental decoding. Taken from (Polák et al., 2023b).

potheses throughout the entire decoding. From the user's perspective, these systems would display either the top hypothesis or a set of hypotheses – critically, a re-translation SST 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 evaluation of the system's latency.

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.

2.3 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; Müller et al., 2016; Bojar et al., 2021). The advantage of the cascade approach is that we can optimize models for each subtask independently. The main advantage of this approach is that ASR and MT tasks have access to larger and more diverse corpora compared to 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). An additional challenge is the *mismatched domains* where MT models trained on written language may not be well-suited to handle spoken language during inference, leading to potential loss in translation quality. 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). Indeed, the loss of prosodic information can have a detrimental effect on translation quality, especially for languages like Slovak, where the same words can be used to express both declarative and interrogative sentences through the use of different melodies. Finally, it's important to note that many languages have no written form, which makes the cascade approach impractical or impossible for such languages (Harrison, 2007).

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., 2021, 2022; Agarwal et al., 2023). In the context of simultaneous speech translation however, both approaches yield competitive performance. For example, in the last year's edition of IWSLT (Anastasopoulos et al., 2022), the best-performing system in two out of three language pairs was our end-to-end model (Polák et al., 2022a). The advantage of the end-to-end model in SST may be due to the fact that it avoids the extra delay caused by ASR-MT collaboration in the cascade (Wang et al., 2022).

In our work, we primarily work with end-to-end models.

3 Modeling Speech

In this section, we briefly[3] introduce some important E2E modeling approaches for speech. The

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

[3] For a more detailed analysis refer to Prabhavalkar et al. (2023).

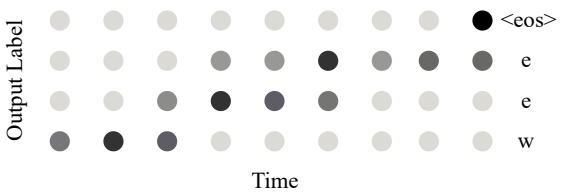

Figure 2: Example of the soft alignment of the word "wee" (small) of an attention-based encoder-decoder (AED) model. The AED model extends the vocabulary with a special symbol <eos> to indicate a completed hypothesis. Each output label c_l (y-axis) attends to all source frames with varying intensity.

purpose of this section is to highlight the differences between the architectures, with an emphasis on how they process the input, and how they align the source speech and the target translation.

We denote the input speech utterance as X. Speech is usually represented as a vector of length T of D-dimensonal acoustic features, i.e., $X = (x_1, \cdots, x_T)$, where $x \in \mathcal{R}^D$ is a feature vector or frame typically representing 10 ms of speech. The corresponding transcript/translation $C = (c_1, \cdots, c_L)$ of length L, consists of a sequence of tokens $c \in \mathcal{C}$, where $\mathcal{C}$ can be characters, words, or other sub-word units. All end-to-end approaches in ASR, ST, and SST use an encoder (e.g., RNN or Transformer, Vaswani et al., 2017) that processes the source X into a vector $H(X) = (h_1, \cdots, h_T)$ of abstract representations. The goal of speech recognition or translation is to model the following conditional distribution:

$$P(C|X) = P(C|H(X)). \tag{1}$$

In the case of the incremental simultaneous ASR/ST, the target of the modeling is the following conditional distribution:

$$P(C_{1:g(X_{1:t})}|X_{1:t}) = P(C_{1:g(X_{1:t})}|H(X_{1:t})), \tag{2}$$

where $t \leq T$ and $g(X_{1:t})$ is a *policy function* that decides how much transcript/translation can be produced given $X_{1:t}$.

3.1 Attention-Based Encoder-Decoder Architecture

In the attention-based encoder-decoder (AED) architecture, the entire source X is first encoded into the abstract representation by an encoder H, and then the attention-based decoder predicts the transcript/translation in a left-to-right autoregressive manner. The alignment is realized by an attention mechanism (Bahdanau et al., 2014) that "attends" to all source positions during each step of generation. Each source position is assigned a score indicating the amount of relevant information coming from that position. Note that many source positions can be relevant to a target token. An example of such an alignment is given in Figure 2.

The AED architecture models the posterior probability as follows:

$$P_{\text{AED}}(C|X) = \prod_{l=1}^{L} P(c_l|a_{l-1}, \cdots, a_0, H(X)), \tag{3}$$

where c_i is an output label at position i, and $c_0 =$ <sos> is a special token that initiates the decoding. The prediction is constructed based on the output length $l \in L$ rather than the source position $t \in T$, a concept we refer to as *output synchrony* (see definition in Section 3.4). Due to this output synchrony, AED incorporates a special token <eos> to indicate a completed hypothesis.

The AED architecture is widely used in offline ST. The advantage is that the decoder can "see" the entire source. This is a particularly important feature for the translation task as it (generally) involves reordering. The fundamental disadvantage of AED is that it cannot work in a simultaneous regime, as it needs to process the entire source first before it starts to produce the translation. To address this, researchers proposed alternatives that guide or prevent the attention to "see" the future source (Raffel et al., 2017; Chiu and Raffel, 2017; Arivazhagan et al., 2019). Another body of work studies chunk-based inference and the "stable hypothesis" selection along with improvements in the beam search decoding (Liu et al., 2020a; Polák et al., 2022a; Polák et al., 2023b).

3.2 Connectionist Temporal Classification

Connectionist temporal classification (CTC, Graves et al., 2006) directly maps the abstract encoder representation of the source directly to labels. To do so, CTC extends the vocabulary $\mathcal{C}$ with a special blank (<b>) label, i.e., $\mathcal{C}_b = \mathcal{C} \cup \{\text{<b>}\}$. In CTC, every source frame is assigned a label, directly modeling both alignment and prediction. An alignment $A \in \mathcal{C}_b^*$ is a string of labels from the vocabulary $\mathcal{C}_b^*$. An example of an alignment is in Figure 3.

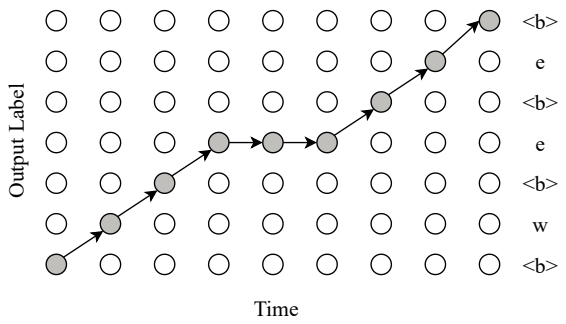

Figure 3: Example of a CTC alignment $A = (\texttt{<b>}, w, \texttt{<b>}, e, e, e, \texttt{<b>}, e, \texttt{<b>})$ predicted by a CTC model for a sequence $C = (w, e, e)$. For each frame, the model outputs a label from the vocabulary $\mathcal{C}$ or a blank. The alignment A is converted to C by collapsing repeated labels and removing blanks. Note that there are multiple possible alignments, and all must be strictly monotonic. I.e, a label c_l can be only emitted after all previous labels $c_1, \cdots, c_{l-1}$ have been emitted.

CTC models the posterior of a label sequence C by marginalizing over all possible CTC alignments $\mathcal{A}^{\text{CTC}}_{(X,C)}$:

$$P_{\text{CTC}}(C|X) = \sum_{A \in \mathcal{A}^{\text{CTC}}_{(X,C)}} P(A|H(X)) = \sum_{A \in \mathcal{A}^{\text{CTC}}_{(X,C)}} \prod_{t=1}^{T} P(a_t|H(X)). \quad (4)$$

As we can see in the second row of Equation (4), CTC makes a strong independence assumption that the label a_t at time t is conditionally independent on all other predictions at $t' \neq t$. Critically, this is a very strong assumption that manifests in lower quality transcripts/translations (Prabhavalkar et al., 2017; Libovický and Helcl, 2018) compared to the AED architecture. However, a combination of CTC and AED architectures showed promising results in offline ASR, MT, and ST (Watanabe et al., 2017; Yan et al., 2023a). In simultaneous regime, Moritz et al. (2019) use CTC output to activate the attention-based decoder for ASR, and Polák et al. (2023a) estimate the position of the AED translation prefix relative to the source.

3.3 Recurrent Neural Transducer

Recurrent neural transducers (RNN-T, Graves, 2012) relax some of the strong independence assumptions in CTC. RNN-T model consists of an encoder H, a predictor R, and a joiner J. Similarly,

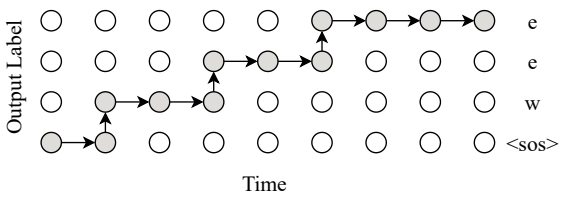

Figure 4: Example of an RNN-T alignment $A = (\texttt{<b>}, w, \texttt{<b>}, \texttt{<b>}, e, \texttt{<b>}, \texttt{<b>}, e, \texttt{<b>}, \texttt{<b>}, \texttt{<b>})$ for sequence $C = (w, e, e)$.

as CTC, RNN-T extends the vocabulary $\mathcal{C}$ with a <b> token, i.e., $\mathcal{C}_b = \mathcal{C} \cup \{\texttt{<b>}\}$. However, unlike CTC, for each source frame, RNN-T predicts zero or more labels $c \in \mathcal{C}$ terminated by exactly one <b> token. In other words, the <b> serves as an indicator that the decoding should move to the next source frame. The set of all possible alignments is then the set of all possible sequences of length $T + L$ in $\mathcal{C}_b^*$: $\mathcal{A}^{\text{RNN-T}}(X, C) = \{A = (a_1, \cdots, a_{T+L})\}$. An example of an RNN-T alignment is in Figure 4.

The following equation defines the posterior probability modeled by RNN-T:

$$P_{\text{RNN-T}}(C|X) = \sum_{A \in \mathcal{A}^{\text{RNN-T}}_{(X,C)}} P(A|H(X)) = \sum_{A \in \mathcal{A}^{\text{RNN-T}}_{(X,C)}} \prod_{\tau=1}^{T+L} P(a_\tau|r_{i_\tau}, h_{\tau - i_\tau}), \quad (5)$$

where i_τ denotes the number of non-blank tokens in the prefix alignment $(a_1, \cdots, a_\tau)$, $r_j = R(r_{j-1}, \cdots, r_0)$ is the output of predictor R conditioned on previously generated non-blank tokens, and h_t is the abstract encoder representation at time t.

The conditional dependence of the predictor R on previously generated tokens makes the RNN-T stronger than CTC. Different from AED architecture, RNN-T still assumes monotonic source-target alignment, which might be a too strong assumption for the translation task. Nevertheless, recent literature suggests that RNN-T might be competitive in ST and SST (Liu et al., 2021; Xue et al., 2022; Tang et al., 2023; Yan et al., 2023b).

3.4 Input- and Output-Synchrony

The previous sections have highlighted a fundamental difference in how the AED processes the source, in contrast to the approaches of CTC and RNN-T.

The AED models follow a two-step process: they begin by encoding the entire utterance and subsequently employ an attention-based decoder to generate the target sequence in an autoregressive left-to-right manner, attending to any desired section of the source. We call this phenomenon an "output synchrony". On the other hand, both CTC and RNN-T produce a monotonic alignment that allows the processing of the source in a left-to-right fashion. We call this an "input synchrony". For monotonic sequence-to-sequence tasks, such as ASR, the input synchrony might be more desirable as it allows for simultaneous decoding, i.e., it naturally follows the gradually arriving input (assuming a unidirectional encoder).

4 Long-form Simultaneous Speech Translation

In the previous section, we discussed various aspects of simultaneous speech translation (SST) in general. 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. In this section, we shift our focus to this real-world long-form setting. We begin by placing long-form SST within the broader context of long-form automatic speech recognition (ASR), machine translation (MT), and offline ST. Subsequently, we explore the current literature on long-form SST.

4.1 Long-Form ASR

In terms of input and output modalities, long-form ASR and ST are facing 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 need to be treated as undivided units. A more sophisticated solution uses CTC blank prediction to indicate non-speech segments (Yoshimura et al., 2020). Another approach, based on RNN-T, performs joint modeling of ASR and sentence segmentation (Huang et al., 2022).

An alternative solution based on fixed segments (Chiu et al., 2019), introduced the concept of overlapping inference for RNN-T models. Here, the utterance is segmented into overlapping (50%) segments. Words from two overlapping segments are merged on the same words. In case of conflicts, predictions further from the boundaries are preferred. Note that this algorithm requires an architecture with explicit source-target alignment. Chiu et al. (2021) extended their previous work and observe that the overlapping inference is particularly important for models with poor generalization to unseen length. The chunking approach was also adopted by the attentional model Whisper (Radford 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 based on LSTM. The findings indicate that only RNN-T and CTC architectures can generalize to unseen lengths. Interestingly, RNN-T in the long-form regime was only 1% worse than their proposed overlapping inference with 16s segments. Narayanan et al. (2019) treat the long-form regime as a domain mismatch problem and explores regularization via training on different domains. Additionally, they suggest simulating context either by randomly sampling the LSTM state from a normal distribution or by passing the state from the previous segment (similar to Dai et al., 2019). Another way how to improve the generalization is to use minimum word error rate training (MWER, Lu et al., 2021).

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 here that while the encoder has a limited attention span, the attention-based decoder sees the entire encoder representation. The model employing AD was unable to function without chunking, whereas the CTC model processed the entire speech at once and still outperformed the AD model.

4.2 Long-Form MT

The primary objective of long-form MT is to enhance textual coherence, as conventional MT systems typically 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). The work showed that MT models benefit from the extra context and handle the inter-sentential discourse phenomena better. 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. In this context, Kim et al. (2019) demonstrated that filtering only essential tokens, such as named entities or words with specific parts-of-speech tags, proved to be beneficial in mitigating the impact of irrelevant information on the model's performance.

Other body of work tries to directly model very long sequences. Dai et al. (2019) introduced a recurrence mechanism and improved positional encoding scheme in the Transformer. The limitation is that the architecture stores previous states in an uncompressed form which increases memory requirements. Later work proposed an explicit compressed memory realized by a few dense vectors (Feng et al., 2022).

4.3 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.[4] The methods are usually based on the re-introduction of 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 they disregard the acoustic information of the source audio. The latter issue was addressed in Iranzo-Sánchez et al. (2020a), however, the approach still requires an intermediate ASR transcript that is not available in E2E models.

Another take on this issue is segmentation based purely on the source speech. The early work focused on segmentation based on VAD. VAD focuses solely on the presence of the speech and disregards sentence boundaries. This usually results in sub-optimal segmentation as humans tend to place pauses inside of sentences and not necessarily between them (e.g., hesitations before words with a high information content, Goldman-Eisler, 1958). To this end, researchers tried to address this by considering not only the presence of speech but also its length (Potapczyk and Przybysz, 2020; Inaguma et al., 2021; Gaido et al., 2021). Later studies tried to avoid VAD and focus on more linguistically-motivated approaches. For example, Gállego et al. (2021) used ASR CTC to predict voiced regions. Further improvements were observed by directly modeling the golden segmentation (Tsiamas et al., 2022b; Fukuda et al., 2022).

To address the problem of inadequate segmentation, Gaido et al. (2020) proposed to leverage previous context and showed that context-aware ST is less prone to segmentation errors. An extensive study of context-aware ST was conducted by Zhang et al. (2021). They compared different chunk-based inference methods, context size, and robustness to segmentation errors. Similarly to long-form MT, they observed that context helps, but this holds only for a limited number of previous utterances in the memory.

4.4 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 that is capable of translation of unsegmented text input. This model could be theoretically adapted for speech input. However, it was later shown that this model exhibits huge latency of up to 100 tokens (Iranzo Sanchez et al., 2022). Another work (Iranzo Sanchez et al., 2022), proposed a partially bidirectional encoder and application of a wait-k policy (Ma et al., 2019) to accommodate the streaming input. They also explored the extended context and confirm the findings from long-form MT and offline ST, demonstrating that the use of the previous context signif-

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

icantly enhances performance. Furthermore, they also confirm that when the context becomes too long, it leads to a drop in translation quality.

Finally, the only direct SST model that claims that it can 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 (typically around 640 milliseconds), to mitigate computational complexity. During the processing, each vector is summarized into a summarization vector and appended to the end of the memory. Unfortunately, based on the reported experiments, it remains unclear whether the model was specifically evaluated in the long-form setting.

4.5 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 Proportion (AP; Cho and Esipova (2016)), differential lagging (DAL; Cherry and Foster (2019)), and average lagging (AL; Ma et al. (2019)). Broadly speaking, they measure "how much of the source was read by the system to translate a word", where the latency unit is typically a word. These metrics were quickly adopted by the speech community. The downside is that the metrics assume a monotonic alignment between the source and the target translation. This issue is further elevated in speech-to-text translation, as the metrics also assume uniform distribution and uniform length of the words in the source. Alternatively, Ansari et al. (2021) proposed to use a statistical word alignment of the candidate translation with the corresponding timestamped source transcript. This theoretically allows for more precise latency evaluation, but it is unclear how the alignment errors impact the reliability of the evaluation.

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, where the reference follows the same segmentation. However, in the long-form unsegmented regime, the candidate and reference segmentation into sentences might differ. Traditionally, this issue was addressed by re-segmenting the hypothesis based on the reference (Matusov et al., 2005). The re-segmentation was done on reference punctuation based on the alignment extracted from a dynamic programming algorithm for edit distance minimization. After the re-segmentation, a typical sentence-level evaluation of translation quality and latency is done. It should be noted that the commonly used latency metrics (AL, AP, 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. To this end, Macháček et al. (2023a) evaluated different translation quality metrics and evaluation setups and their correlation with human judgments. If the candidate and reference segmentation are identical, the results indicate that translation quality exhibits an equal correlation with human judgments at both the sentence and document levels. If the inference and reference segmentation differ, BLEU and COMET correlate significantly more at the document level compared to the sentence level after re-segmentation (Matusov et al., 2005).

In conclusion, we will evaluate the quality using BLEU and COMET on the document level if the candidate and reference segmentation differ. For the latency evaluation, we will use the LAAL (an improved version of AL; (Polák et al., 2022a; Papi et al., 2022)) with re-segmentation, but we carefully check for possible inconsistencies introduced by

the re-segmentation.

5 Goals

In this section, we briefly outline the goals of our research.

Improving the 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. First, we consider the "onlinization", i.e., conversion to the simultaneous regime, of existing state-of-the-art offline ST models. Second, we reconsider the beam search decoding for attention-based encoder-decoder models. Third, we reconsider the simultaneous policy for a joint CTC/AED architecture. Finally, we compare different SST architectures, with special emphasis on time- and label-synchronous architectures, such as CTC, RNN-T, and AED 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. Our goal is to combine segmentation strategies from long-form ASR and long-form ST with existing short-form SST models. Through this study, we aim to investigate the capabilities of end-to-end models to jointly handle translation and segmentation.

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. Our focus will be on identifying an appropriate model architecture and effective training procedures to achieve seamless and reliable long-form simultaneous speech translation.

In the next section, we will describe the results achieved so far.

6 Results

In the first part of our research (presented in Polák et al. (2022a); Polák et al. (2023b,a)), we focused on general improvements in the quality-latency tradeoff. As discussed in Sections 2.3 and 3.1, the AED is very popular in the offline ST. The end-to-end AED architecture makes up the majority of the submissions to the offline track at IWSLT (Anastasopoulos et al., 2022; Agarwal et al., 2023), and is essentially the only alternative to cascade. However, these offline AED models have not typically been used in the simultaneous regime. This is a missed opportunity, especially when considering that these models perform well and are easily trained and available for use. However, in its vanilla form, AED is not capable of simultaneous inference. Therefore, we investigate ways to use offline models in the simultaneous regime. Here, we focus on chunked inference (Liu et al., 2020a; Nguyen et al., 2021), specifically on the stable hypothesis detection. Since we rely on the standard beam search (Sutskever et al., 2014; Bahdanau et al., 2014), the models always generate a complete hypothesis up to the `<eos>` token. Unfortunately, this results in low quality translations, especially towards the end of the hypothesis. Therefore, in Polák et al. (2022a) we investigate methods that select stable hypothesis prefixes that (hopefully) do not contain translation errors. Here, we briefly summarize our findings:

- **We identified the best onlinization technique** (local agreement; Liu et al., 2020a), and proposed varying the chunk size to enable quality-latency tradeoff control.
- We observed that the **models tend to overgenerate**, especially in the low latency regime. This led to a severe quality drop and computation-aware latency increase.
- We **proposed an improved latency metric** based on AL that was robust to overgeneration. This metric was later proposed independently length-adaptive average lagging (LAAL; Papi et al., 2022).
- **Onlinization of offline models is possible.** Across three language pairs (EN $\rightarrow$ DE, JA, ZH) and two models (one trained from scratch and the second based on pre-trained wav2vec 2.0 (Baevski et al., 2020) and mBART (Liu et al., 2020b)) with AL around 2 seconds, translation quality drops only about 0 to 1.5 BLEU compared to the offline baseline.
- **Onlinized offline models are competitive SST models.** In fact, our onlinized model outperformed all other models in EN $\rightarrow$ DE and JA directions and all latency regimes in the simultaneous speech translation track at IWSLT 2022 (Anastasopoulos et al., 2022).

It appears that the local agreement combined with a strong offline model has a competitive performance, but it still suffers from overgeneration and poor translation quality. The root causes are likely to be label/exposure bias (Ranzato et al., 2015; Hannun, 2019) and poor length generalization (Dong et al., 2020; Variš and Bojar, 2021), which manifests itself as hallucinations (Lee et al., 2018). This suggests that the problem should be addressed before the stability detection, i.e., during the decoding. This leads us to revisit the beam search decoding algorithm. In Polák et al. (2023b), we proposed incremental blockwise beam search (IBWBS). We base our algorithm on the blockwise re-translation beam search (BWBS) for the blockwise architecture (Tsunoo et al., 2021), but we adapt it to produce an incremental translation. The key idea of the proposed algorithm is to expand only "reliable" hypotheses. By a "reliable" hypothesis we mean a hypothesis that satisfies all of the following three conditions: (1) it is without <eos> token, (2) without any repeated token,[5] and (3) with a higher score than any unreliable hypothesis. Here is a summary of our findings from Polák et al. (2023b):

- Offline ST models used for SST with the proposed **IBWBS outperformed standard beam search by 5 to 8 BLEU** points and reduced the number of decoder forward passes by 20 %.
- Online blockwise AED/CTC models with **IBWBS outperformed BWBS** by 0.6-3.6 BLEU points in the same latency regime, or **reduced the latency by 0.8-1.4 seconds** with the same translation quality.

While the IBWBS shows significant improvements over the baselines, it still relies on the attention mechanism and the label-synchronous decoding. Ideally, the decoding should be aware of how much of the source has already been covered. This information could prevent the decoding beyond the current context, i.e., hallucination. Vanilla AED models do not produce reliable alignment (see Section 3.1), but current ST models use CTC together with the AED architecture during training and also during inference. As discussed in Section 3.2, CTC produces a monotonic source-target alignment. In Polák et al. (2023a), we explore the potential of CTC to guide the AED decoding. Specifically, we use the CTC prefix probability (Graves, 2008) to estimate the likelihood that the current hypothesis proposed by the AED model covers the current source. Our key findings from Polák et al. (2023a):

- The **proposed CTC policy improves the translation quality** by up to 1 BLEU point compared to our IBWBS in blockwise models.
- For a large offline model, the CTC policy loses up to 1 BLEU point but **reduces the real-time factor to 50 %** compared to our IBWBS.

It is important to note that we failed to train the large models with CTC loss because the mBART decoder vocabulary was too large. Instead, we used a small blockwise model for the CTC posteriors. Since the blockwise model uses a small vocabulary, this resulted in a vocabulary mismatch between the CTC and the AED decoders, which is likely the reason why the large models observed a loss in translation quality of up to 1 BLEU point with the CTC policy. In the future, we would like to explore strategies on how to reduce the memory footprint of CTC with a large vocabulary or how to reduce the vocabulary of large pre-trained models.

7 Future Work

Finally, this leads us to future work. So far, our research has focused primarily on the AED architecture, ignoring other architectures such as CTC and RNN-T. Thanks to the monotonic alignment (discussed in Section 3) and supported by the recent finding for the offline regime (Yan et al., 2023a,b), these architectures have the potential to be advantageous in the simultaneous regime. Our goal will be to compare these architectures for SST. In particular, we will investigate the quality-latency tradeoff, and explore how the monotonicity of CTC and RNN-T affects the translation. We will also evaluate the quality of the alignments and their potential applications.

7.1 Towards the Long-Form SST

As described in the previous section, we already have a solid starting point with capable models for "short-form" SST. But how can we use these models in the more natural long-form scenario?

We can take inspiration from the offline long-form ST, where the main focus is on segmenta-

[5] The repeated token rule is only necessary for the blockwise architecture and is not required for offline models.

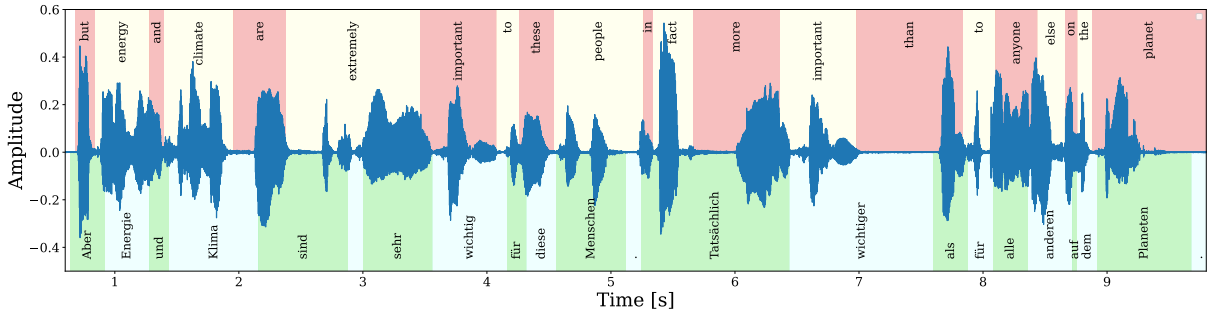

Figure 5: English ASR CTC alignment (top) and English-to-German ST CTC alignment (bottom). The colors represent the timestamps of the aligned word. ASR CTC outputs only lowercase words without punctuation. ST CTC outputs words with correct case and sentence punctuation, including two full stops at 5.2s and 9.8s. The ST CTC alignment closely matches the ASR CTC alignment. In addition, the two full stops are correctly aligned with the silence.

tion. The best approach seems to be direct segmentation modeling classification (Tsiamas et al., 2022a; Fukuda et al., 2022). The limitation of these approaches is that they do not allow simultaneous inference. However, we believe that their adaptation to the simultaneous regime should be relatively straightforward (e.g., by using a unidirectional encoder and a sufficiently small loss in accuracy). A possible improvement could be a multitask translation-segmentation model similar to Huang et al. (2022).

Our hopes go even further and we ask: Can we just train a model to translate and let the model figure out the segmentation? The target side of the translation already contains punctuation marks, so if we also knew the alignment, we could use it to directly segment the utterances directly. Here we turn to our previous work, namely to the CTC policy (see Section 6). The CTC policy already relies on the CTC alignment, and it has shown a very good performance in guiding the simultaneous decoding. Therefore, we will design an experiment to use CTC alignment for simultaneous translation and segmentation. Our unpublished work-in-progress already shows promising results as suggested by some anecdotal evidence in Figure 5. However, our focus will not be limited to CTC alignments only. We will also investigate the RNN-T alignments. We may also consider vanilla AED with whisper-style timestamps.[6]

However, we see another valuable use of the direct speech-to-text 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 in that it virtually eliminates all data without a source transcript, preventing the research community from utilizing potentially valuable data sources. It is 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 alignment evaluation shows promising results, we will explore the feasibility of E2E speech-to-text creation.

Another important research direction is context. As discussed in Sections 4.2 to 4.4, the use of context yields significant improvements in robustness, quality, and overall text coherence. Although this has already been addressed in Zhang et al. (2021), we see a few limitations of this work. First, the authors used the re-translation approach (see Figure 1). We hypothesize that the re-translation approach is less susceptible to the exposure bias problem, i.e., the mismatch between teacher-forced training time and the inference where the previous model's own outputs are available (Ranzato et al., 2015; Hannun, 2019), because it can revise the hypothesis with more speech available. In contrast, the incremental approach must continue with the already generated hypothesis, which may become inconsistent with more speech. Second, the work used reference segmentation.

An additional question is how to accommodate context beyond a few sentences. As pointed out in Sections 4.2 to 4.4, the performance usually drops

[6] https://github.com/linto-ai/whisper-timestamped

with too much context. Some solutions have been suggested (Kim et al., 2019; Feng et al., 2022), but it remains unclear how to adapt them for SST with the specifics of SST in mind (e.g., computational constraints, speech input).

7.2 True Long-Form SST

The ultimate goal of our work is to achieve true long-form simultaneous speech translation. In other words, our goal is to develop a model capable of processing a potentially infinite stream of speech input, without any segmentation or special inference algorithm, and translating it 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 (S)ST (Yan et al., 2023b). Therefore, we will compare the architectures, or possibly their hybrids, in the true long-form regime. In these experiments, we will draw inspiration from related work in ASR and MT (Narayanan et al., 2019; Dai et al., 2019; Feng et al., 2022; Rekesh et al., 2023).

8 Conclusion

In conclusion, this thesis proposal has provided an overview of simultaneous speech translation (SST) and its main challenges, including the quality-latency tradeoff. We have discussed different modeling approaches, focusing on CTC, RNN-T, and attention-based encoder-decoder architectures. Through a comprehensive literature review, we observed the limited research on long-form speech translation. We outlined three main goals of our research with a special focus on long-form speech translation: improving the general quality-latency tradeoff in SST, exploring long-form SST through segmented inference, and ultimately achieving true long-form SST modeling. We have placed these goals in the context of related work and outlined a clear strategy for achieving them. Finally, the feasibility of this thesis is documented by our progress in the SST (Polák et al., 2022a; Polák et al., 2023a,b; Yan et al., 2023b; Agarwal et al., 2023), as well as in other related areas (Kratochvíl et al., 2020; Polák et al., 2020; Polák and Bojar, 2021; Polák et al., 2021; Bojar et al., 2021; Polák et al., 2022b; Macháček et al., 2023b).

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, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019. Monotonic infinite lookback attention for simultaneous machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1313–1323, Florence, Italy. 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.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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.

Chung-Cheng Chiu and Colin Raffel. 2017. Monotonic chunkwise attention. *arXiv preprint arXiv:1712.05382*.

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.

Linhao Dong, Cheng Yi, Jianzong Wang, Shiyu Zhou, Shuang Xu, Xueli Jia, and Bo Xu. 2020. A comparison of label-synchronous and frame-synchronous end-to-end models for speech recognition. *arXiv preprint arXiv:2005.10113*.

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.

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.

Daniel Gile. 2004. Translation research versus interpreting research: Kinship, differences and prospects for partnership. *Translation research and interpreting research: Traditions, gaps and synergies*, 2(1):10–34.

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

Yunchao Gong, Liu Liu, Ming Yang, and Lubomir Bourdev. 2014. Compressing deep convolutional networks using vector quantization. *arXiv preprint arXiv:1412.6115*.

Alex Graves. 2008. *Supervised sequence labelling with recurrent neural networks.* Ph.D. thesis, Technical University Munich.

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.

Awni Hannun. 2019. The label bias problem.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

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.

Jonáš Kratochvíl, Peter Polák, and Ondřej Bojar. 2020. Large corpus of czech parliament plenary hearings. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6363–6367.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation. *openreview.net*.

Jindřich Libovický and Jindřich Helcl. 2018. End-to-end non-autoregressive neural machine translation with connectionist temporal classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3016–3021, Brussels, Belgium. Association for Computational Linguistics.

Dan Liu, Mengge Du, Xiaoxi Li, Ya Li, and Enhong Chen. 2021. Cross attention augmented transducer networks for simultaneous translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 39–55, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Danni Liu, Gerasimos Spanakis, and Jan Niehues. 2020a. Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis Selection. In *Proc. Interspeech 2020*, pages 3620–3624.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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, Ondřej Bojar, and Raj Dabre. 2023a. MT metrics correlate with human ratings of simultaneous speech translation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 169–179,

Toronto, Canada (in-person and online). Association for Computational Linguistics.

Dominik Macháček, Peter Polak, Ondřej Bojar, and Raj Dabre. 2023b. Robustness of multi-source MT to transcription errors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3707–3723, Toronto, Canada. 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.

Niko Moritz, Takaaki Hori, and Jonathan Le Roux. 2019. Triggered attention for end-to-end speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5666–5670. IEEE.

Markus Müller, Thai Son Nguyen, Jan Niehues, Eunah Cho, Bastian Krüger, Thanh-Le Ha, Kevin Kilgour, Matthias Sperber, Mohammed Mediani, Sebastian Stüker, and Alex Waibel. 2016. Lecture translator - speech translation framework for simultaneous lecture translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 82–86, San Diego, California. Association for Computational Linguistics.

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.

Thai-Son Nguyen, Sebastian Stüker, and Alex Waibel. 2021. Super-Human Performance in Online Low-Latency Recognition of Conversational Speech. In *Proc. Interspeech 2021*, pages 1762–1766.

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. 2021. Coarse-to-fine and cross-lingual asr transfer. *arXiv preprint arXiv:2109.00916*.

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. 2022a. 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, Sangeet Sagar, Dominik Macháček, and Ondřej Bojar. 2020. CUNI neural ASR with phoneme-level intermediate step for~Non-Native~SLT at IWSLT 2020. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 191–199, Online. Association for Computational Linguistics.

Peter Polák, Muskaan Singh, and Ondřej Bojar. 2021. Explainable quality estimation: CUNI Eval4NLP submission. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 250–255, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter Polák, Muskaan Singh, Anna Nedoluzhko, and Ondřej Bojar. 2022b. ALIGNMEET: A comprehensive tool for meeting annotation, alignment, and evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1771–1779, Marseille, France. European Language Resources Association.

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. SRPOL'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.

Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe. 2023. End-to-end speech recognition: A survey. *arXiv preprint arXiv:2303.03329*.

Rohit Prabhavalkar, Kanishka Rao, Tara N Sainath, Bo Li, Leif Johnson, and Navdeep Jaitly. 2017. A comparison of sequence-to-sequence models for speech recognition. In *Interspeech*, pages 939–943.

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.

Colin Raffel, Minh-Thang Luong, Peter J Liu, Ron J Weiss, and Douglas Eck. 2017. Online and linear-time attention by enforcing monotonic alignments. In *International conference on machine learning*, pages 2837–2846. 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.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Russell Reed. 1993. Pruning algorithms-a survey. *IEEE transactions on Neural Networks*, 4(5):740–747.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

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.

Yun Tang, Anna Sun, Hirofumi Inaguma, Xinyue Chen, Ning Dong, Xutai Ma, Paden Tomasello, and Juan Pino. 2023. Hybrid transducer and attention based encoder-decoder modeling for speech-to-text tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12441–12455, Toronto, Canada. Association for Computational Linguistics.

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*.

Emiru Tsunoo, Yosuke Kashiwagi, and Shinji Watanabe. 2021. Streaming transformer asr with blockwise synchronous beam search. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 22–29. IEEE.

Dušan Variš and Ondřej Bojar. 2021. Sequence length is a domain: Length-based overfitting in transformer models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8246–8257, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Jian Xue, Peidong Wang, Jinyu Li, Matt Post, and Yashesh Gaur. 2022. Large-Scale Streaming End-to-End Speech Translation with Neural Transducers. In *Proc. Interspeech 2022*, pages 3263–3267.

Brian Yan, Siddharth Dalmia, Yosuke Higuchi, Graham Neubig, Florian Metze, Alan W Black, and Shinji Watanabe. 2023a. CTC alignments improve autoregressive translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1623–1639, Dubrovnik, Croatia. 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. 2023b. 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*.