How ''Real'' is Your Real-Time Simultaneous Speech-to-Text Translation System?

Sara Papi[†] and **Peter Polák**[◇] and **Dominik Macháček**[◇] and **Ondřej Bojar**[◇]
[†]Fondazione Bruno Kessler, Trento, Italy
[◇]Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics, Praha, Czech Republic
[†]spapi@fbk.eu, [◇]{polak, machacek, bojar}@ufal.mff.cuni.cz

Abstract

Simultaneous speech-to-text translation (SimulST) translates source-language speech into target-language text concurrently with the speaker's speech, ensuring low latency for better user comprehension. Despite its intended application to unbounded speech, most research has focused on human pre-segmented speech, simplifying the task and overlooking significant challenges. This narrow focus, coupled with widespread terminological inconsistencies, is limiting the applicability of research outcomes to real-world applications, ultimately hindering progress in the field. Our extensive literature review of 110 papers not only reveals these critical issues in current research but also serves as the foundation for our key contributions. We: 1) define the steps and core components of a SimulST system, proposing a standardized terminology and taxonomy; 2) conduct a thorough analysis of community trends; and 3) offer concrete recommendations and future directions to bridge the gaps in existing literature, from evaluation frameworks to system architectures, for advancing the field towards more realistic and effective SimulST solutions.

1 Introduction

The term ''simultaneous'' was first coined in the field of language interpretation, which is the practice of conveying a speaker's message orally in another language to listeners who would not otherwise understand it.[1] Unlike consecutive interpreting (Paulik and Waibel, 2010; Lv and Liang, 2019), where interpretation occurs after the speaker has finished talking, simultaneous interpreting[2] happens concurrently with the speech.[3]

Applying this concept to computer science, specifically in automatic translation, **simultaneous speech-to-text translation** (SimulST) is defined as the process that ''*translates source-language speech into target-language text concurrently*'' (Ren et al., 2020), meaning that the translation process occurs in parallel with the incremental acquisition of the input speech. Within this context, the **real-time** aspect, i.e., the ''*immediate processing and response to inputs, often within milliseconds to seconds*'' (Laplante, 1992) and, in general, with low latency, is crucial for ensuring the synchronicity between input and output, enhancing user comprehension of the translated content (Bangalore et al., 2012).

In Fügen et al. (2007), the SimulST task has been formalized for the first time and described as the process that takes as input an ''audio stream'', a continuous and unsegmented flow of speech information, and produces the automatic textual translation. Despite this broad definition, the field has since predominantly focused on a much narrower task: translating speech that has been pre-segmented into short utterances of few seconds by humans before translation (Kolss et al., 2008; Cho et al., 2015; Ma et al., 2020b, Zhang et al., 2024, among others), following sentence boundaries. While this approach simplifies the translation process by sidestepping challenges related to audio segmentation (Polák, 2023) and

[1]Source: https://knowledge-centre-interpretation.education.ec.europa.eu/.

[2]It is worth noting that, while this paper draws on the concept of simultaneous human interpreting, which is generally speech-to-speech, our focus here is on speech-to-text translation, with speech-to-speech translation falling outside the scope of this study.

[3]Source: https://www.atanet.org/client-assistance/consecutive-vs-simultaneous-interpreting-whats-the-difference/.

Transactions of the Association for Computational Linguistics, vol. 13, pp. 281–313, 2025. https://doi.org/10.1162/tacl_a_00740
Action Editor: Hermann Ney. Submission batch: 8/2024; Revision batch: 11/2024; Published 4/2025.

selecting audio-textual context to retain from the past (Papi et al., 2024b), it offers an incomplete and overly simplistic view of the broader challenges inherent in translating continuous audio streams.

This narrow focus has been reinforced over the years, and recent surveys have continued to emphasize this view, assuming human-segmented audio as the standard setting for the task (Liu et al., 2024), as well as reinforcing a glaring terminological inconsistency affecting the SimulST literature. Terms such as ''streaming'', ''online'', and ''real-time'' are often used interchangeably with ''simultaneous,'' and many terms are used without explicit definitions, leading to significant ambiguity and confusion in understanding and comparing research work, their results, and subsequent findings, ultimately hindering the progress in the field.

In this paper, we aim to address this *terminological chaos* and provide a clearer understanding of SimulST and all its challenges, with a particular focus on processing continuous audio streams and the difficulties therein. After a brief overview of the speech translation landscape (§2), our contributions are structured as follows:

- We define the steps required to build a SimulST system, from audio acquisition to translation presentation, and propose a unified terminology to standardize the task. We also introduce a taxonomy based on the dichotomies identified in our analysis of fundamental system components (§3).
- We present a comprehensive and systematic survey of 110 relevant papers in the field of SimulST, showing significant terminological inconsistencies in the literature, highlighting the prevalent focus of the research on human-segmented speech, and identifying trends within the community (§4).
- Based on our findings, we advocate for the adoption of coherent terminology in the field and call for a shift in research efforts towards more holistic systems capable of effectively processing and translating continuous audio streams. We also provide general recommendations for the research community and suggest promising directions for future investigations spanning from evaluation frameworks to architectural novelties (§5).

2 Background

2.1 Offline Speech Translation

Offline speech translation (ST) is the task of translating speech from the source language into text in the target language. Differently from simultaneous ST, which processes input incrementally, offline ST deals with complete and typically well-formed speech segments, representing one or more sentences. This task was the first addressed by the community (Waibel, 2004), and its model architectures have evolved significantly over time. Initially, offline ST was tackled using cascade architectures (Stentiford and Steer, 1988; Waibel et al., 1991), consisting of an automatic speech recognition model (ASR) that transcribes the speech content, followed by a machine translation (MT) model that translates the transcript into the target language. Lately, direct architectures—first developed as statistical approaches (Casacuberta et al., 2001; Matusov et al., 2006) and, later, as neural-based models (Bérard et al., 2016; Weiss et al., 2017)—emerged with the promise of overcoming cascade architectures' inherent limitations (Sperber and Paulik, 2020), such as error propagation[4] by bypassing intermediate ASR outputs. Although direct architectures initially faced a performance gap compared to cascade models (Niehues et al., 2018a, 2019), their effectiveness has been steadily improving (Bentivogli et al., 2021), with an increasing number of works adopting this paradigm, as highlighted in the survey by Latif et al. (2023).

2.2 Audio Segmentation

Most contemporary neural systems for speech processing, both cascade and direct models, are primarily designed to handle short utterances due to inherent memory and modeling limitations (Dai et al., 2019; Chiu et al., 2019). To address this, the common approach has been to segment speech into smaller chunks before feeding it into the model. Ever since the early SimulST systems (e.g., Woszczyna et al., 1998; Fügen et al., 2006b, 2007), audio segmentation has been a natural part of the pipeline in practical settings. In cascaded systems,

[4]Errors in the ASR are directly transferred to the MT model, which cannot recover from them, making it more difficult for the user to understand the original content.

a typical method for segmentation involves introducing punctuation into the ASR-generated text[5] (Lu and Ng, 2010; Rangarajan Sridhar et al., 2013; Cho et al., 2015, 2017; Iranzo-Sánchez et al., 2020) and segmenting based on the punctuation obtained for the subsequent steps of the SimulST process. Direct models, which lack an intermediate transcript, rely on segmentation based solely on speech information. Early approaches used voice activity detection (VAD; Sohn et al., 1999), supplemented by some heuristics to improve performance (Potapczyk and Przybysz, 2020; Inaguma et al., 2021; Gaido et al., 2021). Alternatively, fixed-length segmentation, which divides speech into equally sized segments (usually between 10 and 30 seconds), has been found to often outperform VAD-based methods (Sinclair et al., 2014; Gaido et al., 2021). However, both approaches neglect syntactic and semantic cues in speech, leading to suboptimal results for ST (Sinclair et al., 2014; Tsiamas et al., 2022; Polák and Bojar, 2023). To bridge this gap, recent data-driven approaches have been proposed to model sentence-level segmentation (Tsiamas et al., 2022; Fukuda et al., 2022b). Although these methods were initially developed for the offline regime, Gaido et al. (2021) introduced an algorithm that allows them to be applied to SimulST. Despite this advancement, the effectiveness of these methods in simultaneous settings remains limited (Polák and Bojar, 2023).

2.3 Long-Form Speech

Long-form speech refers to long audio segments, such as entire lectures, podcasts, or interviews, where the speech is continuous and unsegmented. In the related field of ASR, handling such inputs typically involves segmenting the audio into smaller segments, commonly using VAD tools to detect pauses or speech boundaries (Atal and Rabiner, 1976; Ferrer et al., 2003; Novitasari et al., 2022). More recent work has introduced approaches where segmentation decisions are embedded directly within the ASR model itself (Yoshimura et al., 2020; Huang et al., 2022). Additionally, some methods employ fixed segmentation with heuristics to stitch segments together, ensuring the continuity of the recognized speech (Chiu et al., 2019; Radford et al., 2023) or explore architectures capable of performing ASR without segmentation, processing the speech in its entirety (Narayanan et al., 2019; Chiu et al., 2019; Lu et al., 2021; Zhang et al., 2023b).

In cascaded ST, the challenge extends to MT systems, which have to handle the long texts generated by ASR models. While segmenting long text is usually guided by punctuation and supported by using past sentences as context (Tiedemann and Scherrer, 2017; Agrawal et al., 2018; Kim et al., 2019; Donato et al., 2021; Fernandes et al., 2021), it becomes challenging when ASR output lacks punctuation. This issue is typically addressed by inserting punctuation (Lu and Ng, 2010; Rangarajan Sridhar et al., 2013; Cho et al., 2017). Recent methods have aimed to completely bypass segmentation, allowing translation models to process continuous text streams and improving translation coherence (Schneider and Waibel, 2020; Iranzo-Sánchez et al., 2024). In direct ST, research on long-form speech has primarily focused on addressing segmentation challenges. Some studies have integrated previous context to improve translation coherence and quality by mitigating audio segmentation errors (Gaido et al., 2020; Zhang et al., 2021; Ahmad et al., 2024). Recent advances in SimulST suggest the potential to completely eliminate external segmentation, significantly reducing latency and improving translation quality (Polák and Bojar, 2023; Papi et al., 2024b).

3 What is Simultaneous Speech-to-Text Translation?

In this section, we present the first contribution of our work. We begin with the definition of steps characterizing the SimulST process (§3.1), and then provide a unified terminology and taxonomy of the current models developed in the field (§3.2).

3.1 Process Decomposition

We describe the SimulST as a 6-step process, deriving it from a high-level conceptualization of the task from which system implementations may depart in many ways. We start with audio acquisition and conclude with the translation presentation to the user. Throughout the paper, we assume the processing of clean non-overlapping speech in one language, delivered by a single speaker. We leave aspects such as robustness to background noise (Chen et al., 2022; Hwang et al., 2024),

[5]Typically, ASR outputs are lowercase words without any punctuation.

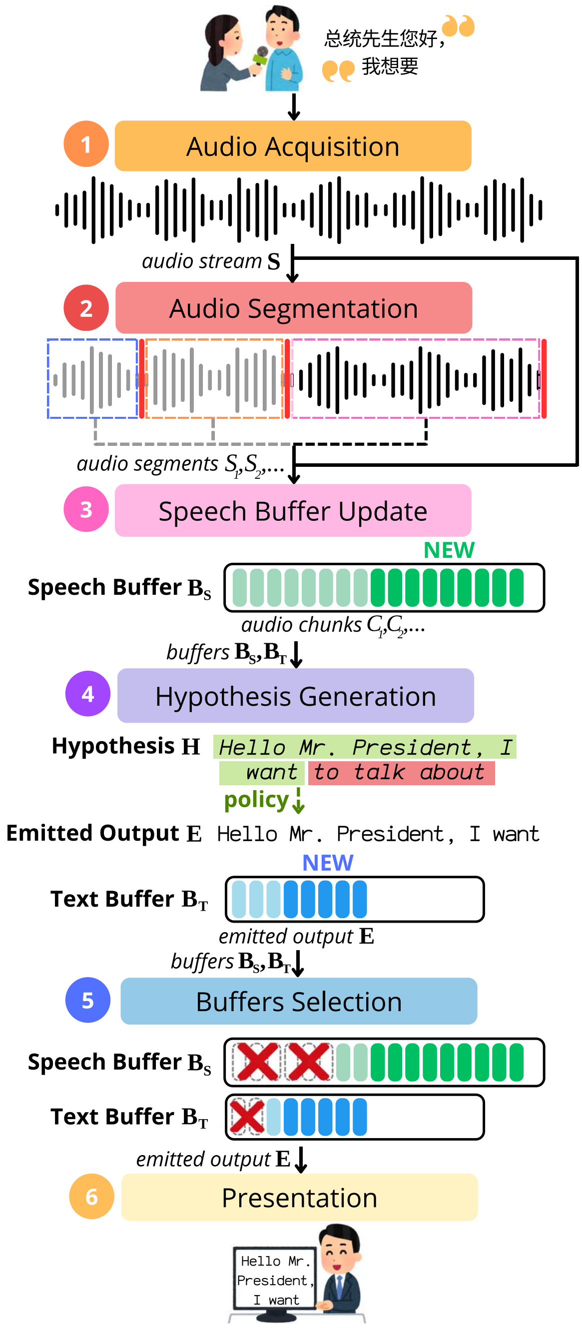

Figure 1: Representation of the steps (1 to 6) of the SimulST process.

speaker diarization (Park et al., 2022), overlapping speech (Wang et al., 2022a), code-switching (Weller et al., 2022; Huber et al., 2022), and any other issues connected to sound to future work on the topic.

The entire process is illustrated in Figure 1 and described as follows:

1. **Audio Acquisition**: The speaker speaks to a microphone that is constantly recording, i.e., collecting the flow of information including unvoiced parts such as pauses or hesitations.

   ➡ **Output:** unbounded speech (audio stream) $\mathbf{S}$.

2. *(Optional)* **Audio Segmentation**: The audio stream $\mathbf{S}$ is segmented into smaller audio segments, usually of a few seconds, based on the utterances contained in the audio using an audio segmenter model.

   ➡ **Output:** bounded speech (audio segments) $\mathbf{S}_{\text{seg}} = [S_1, \dots, S_U]$ where U is the number of utterances detected by the audio segmenter.

3. **Speech Buffer Update**: The incoming speech $\mathbf{S}$ (or current segment $S_u \in \mathbf{S}_{\text{seg}}$) is split into fixed-sized audio chunks (e.g., 500ms each) for incremental feeding to the ST model and the available input is updated. The resulting speech chunks $\mathbf{C} = [C_1, \dots, C_{|\frac{len(\mathbf{S})}{D}|}]$, where D is the fixed-sized duration, are added to the Speech Buffer $\mathbf{B}_\mathbf{S}$, which stores the accumulated speech. The model can process the whole buffer or only part of it at each step.

   ➡ **Output:** The Speech Buffer $\mathbf{B}_\mathbf{S}$ at step t is updated with the new content:

$$\mathbf{B}_\mathbf{S}^\mathbf{t} \leftarrow \mathbf{B}_\mathbf{S}^{\mathbf{t-1}} \oplus \mathbf{C}$$

4. **Hypothesis Generation**: The current Speech Buffer $\mathbf{B}_\mathbf{S}^\mathbf{t}$ is fed into an ST model $\mathcal{M}$ (either cascade or direct) together with the Text Buffer $\mathbf{B}_\mathbf{T}^{\mathbf{t-1}}$, storing the emitted output previously emitted at step $t-1$. The ST model returns the translation hypothesis $\mathbf{H}$:

$$\mathbf{H} \leftarrow \mathcal{M}(\mathbf{B}_\mathbf{S}^\mathbf{t}, \mathbf{B}_\mathbf{T}^{\mathbf{t-1}})$$

   The final output $\mathbf{E}$ is obtained by applying a *decision policy* (Grissom II et al., 2014), which is the strategy determining whether to emit the generated hypothesis or part of it or to wait for more input.

   ➡ **Output:** The new translated text selected by the policy $\mathbf{E} = policy(\mathbf{H})$, which is also appended to the Text Buffer $\mathbf{B}_\mathbf{T}$ at step t:

$$\mathbf{B}_\mathbf{T}^\mathbf{t} \leftarrow \mathbf{B}_\mathbf{T}^{\mathbf{t-1}} \oplus \mathbf{E}$$

5. *(Optional)* **Speech and Text Buffers Trimming**: The content of the Speech and Text Buffers ($\mathbf{B_S}$ and $\mathbf{B_T}$) is trimmed based on the audio-textual information to be retained from the past. This step makes the size of the buffers manageable by ST models, which cannot deal with an infinitely growing context. The content is determined by a *trim* function, which keeps the useful history in the memory for the Hypothesis Generation step (Step 4) at the next step $t+1$:

$$\mathbf{B_S^{t+1}}, \mathbf{B_T^{t+1}} \leftarrow trim(\mathbf{B_S^{t}}, \mathbf{B_T^{t}})$$

   The trim function should ensure semantic alignment of the speech and text buffer contents, as significant misalignment between the two may lead to inaccurate translations by the ST model. If the Audio Segmentation step (Step 2) is applied, both Speech and Text Buffers are typically reset (i.e., completely trimmed $\mathbf{B_{\{S,T\}}^{t+1}} \leftarrow \emptyset$) between each audio segment S_u contained in $\mathbf{S}$.

   ➡ **Output:** The <u>old</u> content contained in the buffers is either reset, trimmed, or left unaltered, providing the Speech and Text Buffers for the next step $\mathbf{B_S^{t+1}}$ and $\mathbf{B_T^{t+1}}$.

6. **Output Presentation**: The translation is either incrementally presented (e.g., word by word, or using meaningful units), or revised (e.g., such as in re-translation).

   ➡ **Output:** The emitted translation $\mathbf{E}$ is displayed to the user.

The SimulST process aims to balance the *quality* and *latency* of spoken content translation, a balance often referred to as the *quality-latency trade-off*. Latency measures the time from when an information is spoken to when the corresponding output is delivered. The quality-latency trade-off is mainly determined by the *decision policy* or, more simply, *policy* in the Hypothesis Generation (Step 4), which decides whether and what part of the hypothesis generated by the model has to be emitted. The decisions made by the policy determine the final output quality and latency, as waiting for more input generally results in higher quality due to increased context but also increases latency. Conversely, emitting output with less context reduces latency but may compromise translation quality.

The Audio Segmentation (Step 2), in which the audio stream is segmented into short utterances, is commonly employed in the SimulST process (see §3.2). This segmentation addresses the current limitations of neural models in processing very long inputs,[6] mainly due to memory constraints (Tay et al., 2022). Utterance boundaries are typically detected using silence-based tools (e.g., VAD, §2.2), but since silence often misaligns with semantic boundaries, newer neural models (e.g., SHAS; Tsiamas et al., 2022) use semantic content for better accuracy, enhancing translation quality. This step is optional for approaches that handle unbounded speech (Polák, 2023; Papi et al., 2024b), where Speech and Text Buffer Trimming (Step 5) becomes crucial to balance past information with the context length manageable by the ST system.

3.2 Terminology and Models' Components

Considering the process described in §3.1, we define the terminology related to the SimulST task in Table 1. This terminology offers a precise and unified framework for understanding and analyzing SimulST models and will be consistently adopted throughout this paper.

Building on this terminology and considering the common distinctions in the context of speech translation (§2), we classify 110 papers proposing SimulST solutions based on their fundamental components, namely: *input* (either bounded or unbounded speech), *architecture* (either direct or cascade), and *output strategy* (either incremental or re-translation). The papers are collected through Semantic Scholar[7] using relevant keywords, whose details and specific categorization are presented in Appendix A. The resulting taxonomy is visualized in Figure 2.

Bounded vs. Unbounded Input Speech. The input of a SimulST system can be either *bounded* or *unbounded* speech, depending on whether the audio has been pre-segmented into sentences in advance (i.e., offline) or not. Bounded speech refers to short audio segments, usually of a few seconds, representing one or more sentences,[8]

[6] Suffice it to say that audio input is at least one order of magnitude longer than textual input.

[7] `https://www.semanticscholar.org/`.

[8] Sentence-level segmentation should not be confused with word-level segmentation, which is commonly used in SimulST policies (Ma et al., 2020b; Dong et al., 2022; Zhang and Feng, 2023) to determine which words to emit.

Term	Definition
simultaneous	concurrently receiving input and generating output
real-time	processing and response to inputs with low latency
policy	the rules regulating when to emit output versus when to wait for more input
incremental	sequential over time rather than all at once
re-translation	the process of generating hypothesis and revising (either entirely or partially) the previously emitted translation
unbounded	a long stream without any explicit information about the overall length
bounded	inputs with a limited length
segmentation	the process that splits unbounded inputs into bounded inputs
segmentation-free	an approach that works on unbounded inputs and does not require segmentation
pre-segmentation	the segmentation is applied to the input before starting the translation process
audio stream	a continuous and unsegmented flow of speech data
audio segment	a portion of speech of a few seconds resulting from the audio segmentation process
audio chunk	a short piece of audio information, usually of fixed length (e.g., $500ms$), used for incremental feeding into ST models
computationally unaware latency	a metric that measures the time between when information is spoken to when the corresponding output is delivered, assuming zero model computation time
computationally aware latency	a metric that measures the time from when information is spoken to when the corresponding output is delivered, also accounting for the model's actual computation time

Table 1: Proposed terminology for the SimulST task.

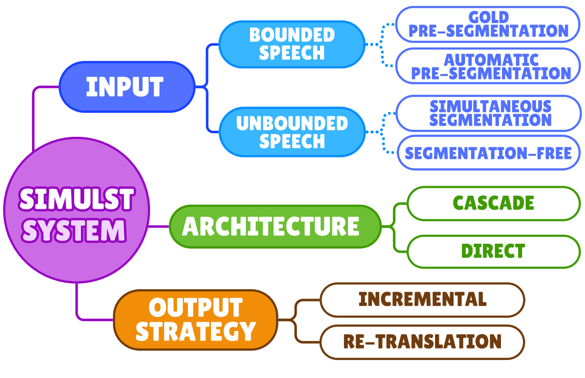

Figure 2: Taxonomy of the SimulST solutions.

while unbounded speech refers to long audio segments or streams with an unknown duration (§2.3). When the input is unbounded and the system processes audio streams directly without any segmentation step (without Step 2 in Section 3.1), we categorize it as a *segmentation-free* system (Iranzo-Sánchez et al., 2024). In this case, selecting the speech and text history to retain from the past—stored in the Speech and Text Buffers (Step 5 in §3.1)—is crucial since audio streams do not have a clear beginning and end, leading to a growing audio-textual context without an explicit resetting mechanism (Polák et al., 2023; Papi et al., 2024b). When the input is unbounded but the system integrates an audio segmentation mechanism that operates jointly with the model in real-time (Step 2 in §3.1), we use the term *simultaneous segmentation* (Fügen et al., 2007). In this case, the history to retain from the past is reset between each automatically detected audio segment. When the input is bounded, the system is not responsible for audio segmentation or managing the growing context of processing incremental audio streams. Instead, it only handles the hypothesis generation (Step 4, §3.1), starting from either *automatically pre-segmented audio* (e.g., using VAD tools) or *gold pre-segmented speech* (i.e., audio manually split or post-edited by humans).

Direct vs. Cascade Architecture. Direct or end-to-end ST architectures are systems that "*translate speech without using explicitly generated intermediate ASR output*" (Sperber and Paulik, 2020). This definition extends to the simultaneous translation scenario, distinguishing direct approaches from cascade architectures that employ separate ASR and MT systems, where the best hypothesis of the former serves as input to the latter. Bahar et al. (2019) surveyed various direct architectures, many of which leverage multi-task training (Luong et al., 2016)—e.g., incorporating Connectionist Temporal Classification (CTC) loss computed on transcripts (Graves et al., 2006) alongside standard cross-entropy loss—and pre-training techniques (Bansal et al., 2018, 2019)—e.g., initially training on the ASR task before the ST task—to enhance model performance. In the context of simultaneous translation, the most prevalent direct architectures include single-encoder single-decoder models (e.g., Ma et al., 2020b), double-encoder models (e.g., Chen et al., 2021), and double-decoder models (e.g., Ren et al., 2020; Zeng et al., 2021).

Incremental vs. Re-translation. SimulST systems produce partial translations to provide a real-time experience to the end user. Based on their output strategies, these systems are categorized into *incremental* and *re-translation*. Re-translation (Niehues et al., 2016, 2018b) allows the system to revise its previous outputs, even after they have been shown to the user. Each time, the SimulST system generates the best translation based on the current incremental speech input and decides whether to change the previous partial translation, either entirely or partially (Chen et al., 2023). The advantage of this approach is that the final translation can achieve a comparable translation quality to an offline system (Arivazhagan et al., 2020a). However, frequent changes in the translation can be challenging to process for users, as they need to identify and re-read the updated parts of the translation (Arivazhagan et al., 2020b), causing many saccades (i.e., quick movements of eyes). Consequently, evaluating the stability of the emitted output and the flickering phenomena (i.e., how frequently the visualized output changes and how far back the user has to scan to see updates), referred to as *stability-latency trade-off* (Arkhangorodsky et al., 2023), has become an integral part of re-translation system assessment (Zheng et al., 2020). Differently, incremental systems (Cho and Esipova, 2016; Dalvi et al., 2018) update the translation shown to the user only by appending new tokens. While a wrong output cannot be corrected in subsequent steps, this approach ensures complete stability of the output, minimizing user cognitive effort and eye movements due to the absence of revisions in the visualized output (Gegenfurtner, 2016). Moreover, incremental systems are also well-suited for speech output, where the produced sound can only be extended and never revised.

Computationally Aware vs. Unaware Latency. The output of a SimulST system is typically evaluated in terms of both quality and latency, as already mentioned in §3.1. Latency metrics can be computed in two ways based on how time-stamps are assigned to each emitted word or character: either by assuming the *ideal* time, i.e., with zero computational overhead, referred to as *computationally unaware latency*, or by considering the actual *elapsed* time of producing the output, known as *computationally aware latency* (Ma et al., 2020a). Unlike the computationally unaware latency, which captures aspects such as the timing of decisions made by the SimulST policy and differences in word order between languages, the computationally aware latency includes both the computationally unaware latency and the actual computational time required for the entire process. This measure provides a more realistic assessment of the latency of the SimulST system (Ma et al., 2020b), but it is strongly influenced by external factors such as the hardware and process optimization being applied (e.g., a more efficient codebase).

4 Is it ‘‘Real’’ Simultaneous Translation?

In the following, we analyze and discuss the results obtained by categorizing the papers using the taxonomy depicted in Figure 2 and whose differences are discussed in §3.2.

The Terminological Chaos. Although ‘‘simultaneous’’ is the most widely adopted term by the research community to refer to the concurrent speech-to-text translation task, mentioned in 100 out of 110 papers, it is not the only term used in the literature. Other commonly used synonyms include ‘‘streaming’’, ‘‘online’’, and ‘‘real-time’’. While ‘‘streaming’’ is tied to ASR research, where it indicates a model capable of processing incremental speech inputs with the lowest latency possible (Zhang et al., 2020; Moritz et al., 2020), ‘‘online’’ serves to describe the SimulST task as a counterpart to offline speech translation (Ansari et al., 2020; Anastasopoulos et al., 2021, 2022; Agarwal et al., 2023). Instead, ‘‘real-time’’ is frequently misused to indicate a process that guarantees low latency, which is a goal rather than an accurate description of the concurrent translation task itself. We visualize this terminological chaos in Figure 3, which shows that over 65% of the papers mix and match these terms. Specifically, 39

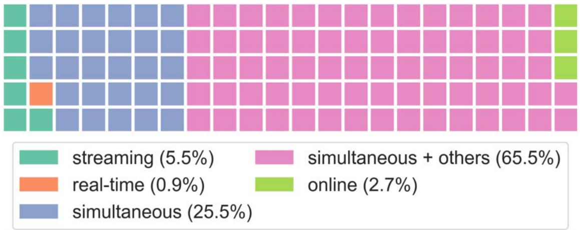

Figure 3: Waffle plot of the term ''simultaneous'' and commonly used synonyms (''streaming'', ''real-time'', and ''online'') among the 110 categorized papers.

papers use at least one of ''streaming'', ''online'', or ''real-time'' terms (mostly opting for the former two) interchangeably with ''simultaneous'' within the same document, 30 papers employ two of the synonyms (preferring ''streaming'' and ''online'' over other combinations), and 3 papers even use all four terms. Moreover, some papers exclusively use ''real-time'' (1 paper) or ''streaming'' (6 papers) to denote the simultaneous translation task, further adding to the confusion. This inconsistent terminology creates significant ambiguity, making it challenging to understand the tasks being addressed, especially when terms are used without explicit definitions. The lack of uniformity calls for a clear, consistent, and standardized task definition in the research landscape, which we addressed in §3.2.

Humans Will Not Segment Our Audio. Despite the inherent complexity of SimulST, only a few works address the task from the beginning by handling unbounded speech inputs (§3.1). Specifically, only 20 papers out of 110 either tackle the concurrent audio segmentation problem for the simultaneous scenario (14 papers) or directly deal with audio streams using a segmentation-free approach (6 papers). In stark contrast, most papers (up to 81.8%) rely on pre-segmented audio as input to their simultaneous models, with nearly all of them (97.7%) using gold segmentation. This approach oversimplifies the real-world scenario where simultaneous translation is performed, as it is impractical to expect human intervention to segment incoming audio before it is fed to the system. Although simplifying assumptions are common in research, an astonishing 91.8% of the papers do not explicitly acknowledge that they assume gold pre-segmented speech for their work. This oversight means that the majority of research bypasses the challenges associated with simultaneous audio segmentation or with the infinitely growing input, as discussed in §3.2, and silently focuses on the optimal hypothesis generation (Step 4, §3.1). Moreover, when examining the bounded speech scenario further, we found only 2 papers (Kolss et al., 2008; Shimizu et al., 2013) that explore the impact of substituting gold segmentation with automatic segmentation. Consequently, our analysis highlights how divisive the issue of processing unbounded speech is within SimulST research: a small fraction of research efforts comprehensively analyze and propose solutions for the entire process, while the majority largely ignores these aspects, operating under unrealistic assumptions that are also rarely explicitly mentioned.

A Clear Trend: Direct Models and Incremental Output. Direct models have quickly gained dominance in the SimulST task due to their potential to decrease latency compared to cascade architectures (Anastasopoulos et al., 2022). Among the 110 categorized papers, 64 versus 49 opted for a direct architecture to address the task. This is even more pronounced in the bounded speech scenario, where 67.8% of the papers leverage a direct approach while being a relatively unaddressed topic in the unbounded speech scenario, with only 3 out of 20 papers using a direct model in their backbone. This trend is also clear in Figure 4, which shows that, since their introduction, an increasing number of work employed direct architectures, almost triplicating from 2021 to 2023, while the number of cascade architectures is steadily decreasing after 2020. The preference for direct models is complemented by a clear prevalence of the incremental output strategy, with 93 out of 110 papers adopting it. Interestingly, in the subset of papers adopting the re-translation strategy, cascade architectures emerge as the preferred choice, with 9 out of 13 papers opting for them. This preference for cascade models in re-translation scenarios contrasts with the general trend in SimulST research, where direct models coupled with incremental output strategies are favored.

5 Recommendations and Future Directions

In this section, we outline best practices derived from the analysis in §4 and the recent advances in the field (⚠️), and we highlight key areas where

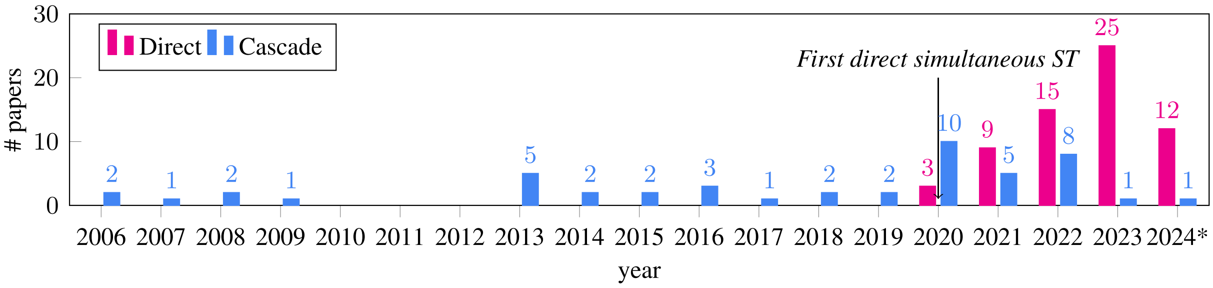

Figure 4: Number of papers in our survey employing direct or cascade simultaneous ST architectures throughout the years. 2024* means that the data are incomplete since the year was not over yet.

future research is needed to develop more robust, accurate, and efficient SimulST systems capable of meeting real-world demands (💡).

⚠️ Use (at least) Automatic Pre-Segmentation. As discussed in §4, the SimulST community has predominantly relied on using gold segmentation for training and evaluating their systems. Since this represents unrealistic conditions for real-world SimulST applications, we encourage future research in the bounded speech scenario to use automatic segmentation instead as input for their models. Offline automatic audio segmentation can be achieved using VAD or neural-based tools such as SHAS (§2.2). Although all audio files are segmented before starting the simultaneous process, they provide a more realistic input, closer to real-world scenarios where audio segmentation (if any) is performed automatically and on the fly. This shift will better prepare models for practical deployment, ensuring that they can handle the challenges of processing speech that is not always segmented into well-formed sentences.

⚠️ Be Clear about the Type of Speech Input. While it may sound like a trivial recommendation, it turns out that a vast majority of papers currently neglect the input conditions specification on which the proposed systems work (as highlighted in §4). Most SimulST research assumes gold segmentation as the default input for their models, implying that the input is bounded and offline pre-segmented (in advance), a condition that has to be explicitly stated in the experimental settings but almost never is. Some papers only detail the size of the speech chunks that are fed incrementally to the model, which, however, alone does not define the type of speech input but only describes how the information is transferred to the model. Explicitly stating the input type (e.g., gold pre-segmented bounded speech) will provide a more accurate understanding of what are the challenges faced by these systems in practice and has to be included in the model description or, at least, in the experimental settings.

⚠️ Always Report Computationally Unaware Latency (and Optionally Aware). Latency is one of the key criteria used to evaluate SimulST systems (§3.1), and all papers report at least one latency metric. However, there is some variation in how these metrics are presented: Some papers report only theoretical (or computationally unaware) latency, others report only computationally aware latency, and a few provide both. Furthermore, in papers using computationally aware metrics, the values are sometimes taken from prior works without recalculating them, even though these metrics are irreproducible without the same hardware setup (§3.2). Given these challenges, we suggest that all papers report computationally unaware metrics, which are always comparable across different hardware setups since they rely solely on theoretical measures. When feasible, computationally aware latency should also be reported, as it provides insight into the real-time usability of the proposed SimulST system, especially when complex or large architectures are involved. In such cases, it is essential to use the same environment (e.g., the GPU and CPU used for running the models and, possibly, the same codebase), for collecting time measurements of the different models being compared to ensure consistency in the resulting metrics.

💡 Create an Evaluation Framework for Unbounded Speech. The most widely adopted evaluation framework for SimulST is SimulEval (Ma et al., 2020a), with 61 out of 110 papers using the tool, which integrates popular metrics for assessing model performance in terms of both quality (e.g., BLEU; Papineni et al., 2002), and latency (e.g., AL: Ma et al., 2019; DAL: Cherry and Foster, 2019; LAAL: Polák et al., 2022, Papi et al., 2022b; and ATD: Kano et al., 2023). However, SimulEval and the aforementioned latency and quality metrics are not designed to compute scores for audio streams and primarily rely on gold pre-segmented inputs. As a result, researchers addressing unbounded speech scenarios have proposed theoretical extensions to these metrics (e.g., StreamLAAL: Papi et al., 2024b) but have resorted to bounded speech scenarios anyway for comparisons (Polák et al., 2023; Papi et al., 2024b). This involves calculating sentence-level scores on automatically aligned audio segments adopting tools such as mWERSegmenter (Matusov et al., 2005), which is commonly used in ST to handle different audio segmentations between reference and output (Anastasopoulos et al., 2021, 2022; Agarwal et al., 2023). However, mWERSegmenter is prone to alignment errors, which complicates the reliability of the evaluation. These reliability issues also impact SLTev (Ansari et al., 2021), another tool for SimulST model assessment. Despite including useful additions such as stability metrics for re-translation and neural-based quality metrics (e.g., COMET: Rei et al., 2020, 2022), SLTev still relies on automatic re-alignment. Another promising starting point is the more recent framework proposed by Huber et al. (2023), which, however, is not as user-friendly as SimulEval, again relies on mWERSegmenter for the alignment, and is currently scarcely adopted.[9] Given the limitations of the current frameworks and metrics, there emerges a clear need for easy-to-use evaluation methodologies and tools also tailored to the more realistic use case of unbounded speech. Such tools should integrate document-level metrics (e.g., as in SLTev) instead of only sentence-level scores, enabling comparisons between systems that handle audio streams without relying on artificial segmentation settings. This advancement would represent an important step towards shifting the community focus on the unbounded speech scenario, more accurately reflecting the real-world conditions in which SimulST systems operate.

💡 **Bear in Mind the Context when Translating.** Real-world applications of SimulST require systems to operate continuously, processing unbounded speech for extended periods. In such scenarios, the context received so far is a valuable source of information that can be employed to improve the accuracy of the provided translations. Despite its significance, research explicitly addressing this aspect in SimulST remains limited. Existing studies explored the use of memory banks to store relevant information (Wu et al., 2020), but these solutions are either not suitable for the unbounded speech scenario (Raffel and Chen, 2023) or claim to support unbounded speech without providing empirical evidence (Ma et al., 2021). Beyond SimulST, a limited number of studies focused on explicitly providing context to the ST model for enhancing translation accuracy. Previous approaches include jointly performing document- and sentence-level translation (Zhang et al., 2021) or integrating context through mechanisms like cross-attention (Gaido et al., 2020). The selection and memorization of the most relevant information during the translation process is an aspect of particular interest for future research, especially in relation to the emerging paradigm of integrating speech foundation models and large language models for addressing a wide variety of tasks (Latif et al., 2023), including speech translation (Gaido et al., 2024), where elements such as prompts and in-context learning (Brown et al., 2020) become of fundamental importance.

💡 **Pay Attention to Output Visualization.** An important factor impacting user experience is how the output is delivered. For textual content such as translations, this primarily concerns how they are visualized on the screen (Romero-Fresco, 2011). Little work has been devoted to this aspect and existing studies have framed the generated texts as subtitles (Machácek and Bojar, 2020; Irvin, 2021; Javorský et al., 2022) and proposed subtitle-oriented metrics (Papi et al., 2021), such as reading speed (Perego et al., 2010), to measure user effort. The aforementioned work also discussed various strategies for delivering the output based on subtitle granularity (i.e., word, lines, and subtitle blocks). However, few studies (Javorský et al., 2022) have examined the impact of SimulST visualization strategies on user comprehension of the generated content or the cognitive effort introduced by translation revisions (§3.2). For instance, the flickering effect inherent to re-translation approaches (Arivazhagan et al.,

[9] At the time of writing, this tool is not even available at the link provided in the paper.

2020b) can cause poor user experience due to re-reading phenomena (Rajendran et al., 2013) and excessive eye fixations (Romero-Fresco, 2010). Therefore, an important future direction for the field is to quantify the effect of output visualization on user comprehension, for instance, by involving human evaluation. Moreover, segmenting the translations for visualization purposes can potentially lead to an overall increased latency of the SimulST systems due to the added processing module. Current subtitle segmentation models, which insert line breaks to satisfy syntactic and semantic constraints for improved readability, were mainly developed for offline ST and are not optimized for low latency or to deal with limited context (Matusov et al., 2019; Karakanta et al., 2020). An alternative approach proposed by Papi et al. (2022c) integrates segmentation directly into the sequence-to-sequence model, potentially reducing latency by bypassing additional modules, and represents an interesting direction for further research.

Quantify Quality-Latency Differences in User Experience. The main goal of SimulST research is to maximize translation quality while minimizing latency, aiming for the best quality-latency trade-off. However, few studies have examined the extent to which variations in quality and latency—whether minor or significant—actually impact user experience (Irvin, 2021; Fantinuoli and Wang, 2024), as well as how automatic translations compare to human interpretations (Bizzoni et al., 2020; Fantinuoli and Prandi, 2021). Assessing and scoring different SimulST systems with humans in the loop remains a challenging area of ongoing research (Sakamoto et al., 2013), as existing methods often suffer from low agreement between participants (Fantinuoli and Wang, 2024). Javorský et al. (2022) proposed and analyzed the effects of continuous ratings (where human evaluators watch videos or listen to audio with translations created by the model being evaluated and continuously express satisfaction by pressing buttons) against traditional questionnaires, but only for re-translation systems. Later, the continuous rating was shown to correlate with standard quality metrics (Macháček et al., 2023), but its generalizability across different domains and systems remains uncertain. Future studies should focus not only on ranking different systems but also on providing holistic human judgments for SimulST outputs, placing the user at the center of the evaluation. Quantifying the minimum changes in the quality-latency trade-off that humans can perceive is of the utmost importance to ensure that improvements measured with automatic metrics also have a meaningful impact on final performance.[10]

6 Conclusions

In this paper, we examined the state of simultaneous speech translation research under several aspects, identifying significant gaps in the existing literature. Our analysis of 110 papers revealed a predominant focus in SimulST on human-segmented speech, which oversimplifies the task and neglects the complexities of real-world applications. We also uncovered substantial terminological inconsistencies, revealing real terminological chaos. To address these issues, we formalized the SimulST task as a 6-step process and introduced a unified terminology to standardize research outcomes. We identified the core components of SimulST systems (input, architecture, and output strategy), discussed current research trends, and provided key recommendations, including transitioning from human to automatic segmentation and adopting consistent terminology. We also emphasized the need for improvement in current evaluation frameworks, highlighting the importance of creating an easy-to-use tool that can handle unbounded speech, incorporating contextual information during translation, and investigating more user-centric assessments to ensure that improvements measured by automatic metrics align with those in the user experience.

Acknowledgments

This paper has received funding from the European Union's Horizon research and innovation programme under grant agreement no. 101135798, project Meetween (My Personal AI Mediator for Virtual MEETings BetWEEN People), from the Ministry of Education, Youth and Sports of the Czech Republic Project nr. LM2023062 LINDAT/CLARIAH-CZ and Project OP JAK Mezisektorová spolupráce Nr. CZ.02.01.01/00/23_020/0008518 named

[10] Refer to Kocmi et al. (2024) for a study of meaningful score differences for MT metrics.

''Jazykověda, umělá inteligence a jazykové a řečové technologie: od vázkumu k aplikacím.'' The authors also acknowledge the support of National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO.

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. `https://doi.org/10.18653/v1/2023.iwslt-1.1`

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.

Ibrahim Said Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, Barry Haddow, Dávid Javorský, Mateusz Krubiński, Tsz Kim Lam, Xutai Ma, Prashant Mathur, Evgeny Matusov, Chandresh Maurya, John McCrae, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, Atul Kr. Ojha, John Ortega, Sara Papi, Peter Polák, Adam Pospíšil, Pavel Pecina, Elizabeth Salesky, Nivedita Sethiya, Balaram Sarkar, Jiatong Shi, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Alex Waibel, Shinji Watanabe, Patrick Wilken, Petr Zemánek, and Rodolfo Zevallos. 2024. Findings of the IWSLT 2024 evaluation campaign. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11, Bangkok, Thailand (in-person and online). Association for Computational Linguistics. `https://doi.org/10.18653/v1/2024.iwslt-1.1`

Belen Alastruey, Matthias Sperber, Christian Gollan, Dominic Telaar, Tim Ng, and Aashish Agarwal. 2023. Towards real-world streaming speech translation for code-switched speech. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 14–22, Singapore. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2023.calcs-1.2`

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. https://doi.org/10.18653/v1/2022.iwslt-1.10

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. https://doi.org/10.18653/v1/2021.iwslt-1.1

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. https://doi.org/10.18653/v1/2020.iwslt-1.1

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. https://doi.org/10.18653/v1/2021.eacl-demos.9

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. 2020a. 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. https://doi.org/10.18653/v1/2020.iwslt-1.27

Naveen Arivazhagan, Colin Cherry, I. Te, Wolfgang Macherey, Pallavi Baljekar, and George Foster. 2020b. Re-translation strategies for long form, simultaneous, spoken language translation. In *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7919–7923. https://doi.org/10.1109/ICASSP40776.2020.9054585

Arkady Arkhangorodsky, Christopher Chu, Scot Fang, Denglin Jiang, Yiqi Huang, Ajay Nagesh, Boliang Zhang, and Kevin Knight. 2023. Method and system for evaluating and improving live translation captioning systems. US Patent US20230089902A1.

Bishnu Atal and Lawrence Rabiner. 1976. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(3):201–212. https://doi.org/10.1109/TASSP.1976.1162800

Parnia Bahar, Tobias Bieschke, and Hermann Ney. 2019. A comparative study on end-to-end speech to text translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 792–799. IEEE. https://doi.org/10.1109/ASRU46091.2019.9003774

Parnia Bahar, Patrick Wilken, Tamer Alkhouli, Andreas Guta, Pavel Golik, Evgeny Matusov, and Christian Herold. 2020. Start-before-end and end-to-end: Neural speech translation by AppTek and RWTH Aachen University. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 44–54, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.iwslt-1.3

Parnia Bahar, Patrick Wilken, Mattia A. Di Gangi, and Evgeny Matusov. 2021. Without further ado: Direct and simultaneous speech translation by AppTek in 2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 52–63, Bangkok, Thailand (online). Association for

Computational Linguistics. https://doi.org/10.18653/v1/2021.iwslt-1.5

Srinivas Bangalore, Vivek Kumar Rangarajan Sridhar, Prakash Kolan, Ladan Golipour, and Aura Jimenez. 2012. Real-time incremental speech-to-speech translation of dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 437–445, Montréal, Canada. Association for Computational Linguistics.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018. Low-resource speech-to-text translation. In *Proceedings of Interspeech 2018*, pages 1298–1302. https://doi.org/10.21437/Interspeech.2018-1326

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 58–68, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1006

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussá, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

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. https://doi.org/10.18653/v1/2021.acl-long.224

Alexandre Bérard, Olivier Pietquin, Laurent Besacier, and Christophe Servan. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*.

Yuri Bizzoni, Tom S. Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? Comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.iwslt-1.34

Ondřej Bojar, Vojtěch Srdečný, Rishu Kumar, Otakar Smrž, Felix Schneider, Barry Haddow, Phil Williams, and Chiara Canton. 2021. Operating a complex SLT system with speakers and human interpreters. In *Proceedings of the 1st Workshop on Automatic Spoken Language Translation in Real-World Settings (ASLTRW)*, pages 23–34, Virtual. Association for Machine Translation in the Americas.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz

Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Francisco Casacuberta, David Llorens, Carlos Martinez, Sirko Molau, Francisco Nevado, Hermann Ney, Moisés Pastor, David Pico, Alberto Sanchis, Enrique Vidal, and Juan M. Vilar. 2001. Speech-to-speech translation based on finite-state transducers. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 1, pages 613–616 vol.1. https://doi.org/10.1109/ICASSP.2001.940906

Chih-Chiang Chang and Hung yi Lee. 2022. Exploring continuous integrate-and-fire for adaptive simultaneous speech translation. In *Proceedings Interspeech 2022*, pages 5175–5179. https://doi.org/10.21437/Interspeech.2022-10627

Chen Chen, Nana Hou, Yuchen Hu, Shashank Shirol, and Eng Siong Chng. 2022. Noise-robust speech recognition with 10 minutes unparalleled in-domain data. In *ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4298–4302. https://doi.org/10.1109/ICASSP43922.2022.9747755

Junkun Chen, Mingbo Ma, Renjie Zheng, and Liang Huang. 2021. Direct simultaneous speech-to-text translation assisted by synchronized streaming ASR. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4618–4624, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.406

Junkun Chen, Jian Xue, Peidong Wang, Jing Pan, and Jinyu Li. 2023. Improving stability in simultaneous speech translation: A revision-controllable decoding approach. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. https://doi.org/10.1109/ASRU57964.2023.10389709

Xinjie Chen, Kai Fan, Wei Luo, Linlin Zhang, Libo Zhao, Xinggao Liu, and Zhongqiang Huang. 2024. Divergence-guided simultaneous speech translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17799–17807. https://doi.org/10.1609/aaai.v38i16.29733

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, Rohit Prabhavalkar, Zhifeng Chen, Tara Sainath, and Yonghui Wu. 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. https://doi.org/10.1109/ASRU46091.2019.9003854

Eunah Cho, Christian Fügen, Teresa Hermann, Kevin Kilgour, Mohammed Mediani, Christian Mohr, Jan Niehues, Kay Rottmann, Christian Saam, Sebastian Stüker, and Alex Waibel. 2013. A real-world system for simultaneous translation of German lectures. In *Proceedings of Interspeech 2013*, pages 3473–3477. https://doi.org/10.21437/Interspeech.2013-612

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, Da Nang, Vietnam.

Eunah Cho, Jan Niehues, and Alex Waibel. 2017. NMT-based segmentation and punctuation insertion for real-time spoken language translation. In *Proceedings Interspeech 2017*, pages 2645–2649. https://doi.org/10.21437/Interspeech.2017-1320

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. https://doi.org/10.18653/v1/P19-1285

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. https://doi.org/10.18653/v1/N18-2079

Keqi Deng, Shinji Watanabe, Jiatong Shi, and Siddhant Arora. 2022. Blockwise streaming transformer for spoken language understanding and simultaneous speech translation. In *Proceedings of Interspeech 2022*, pages 1746–1750. https://doi.org/10.21437/Interspeech.2022-933

Keqi Deng and Phil Woodland. 2024. Label-synchronous neural transducer for E2E simultaneous speech translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8235–8251, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.448

Florian Dessloch, Thanh-Le Ha, Markus Müller, Jan Niehues, Thai-Son Nguyen, Ngoc-Quan Pham, Elizabeth Salesky, Matthias Sperber, Sebastian Stüker, Thomas Zenkel, and Alexander Waibel. 2018. KIT lecture translator: Multilingual speech translation with one-shot learning. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 89–93, Santa Fe, New Mexico. 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. https://doi.org/10.18653/v1/2021.acl-long.104

Qian Dong, Yaoming Zhu, Mingxuan Wang, and Lei Li. 2022. Learning when to translate for streaming speech. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 680–694, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.50

Maha Elbayad, Laurent Besacier, and Jakob Verbeek. 2020a. Efficient wait-k models for simultaneous machine translation. In *Proceedings of Interspeech 2020*, pages 1461–1465. https://doi.org/10.21437/Interspeech.2020-1241

Maha Elbayad, Ha Nguyen, Fethi Bougares, Natalia Tomashenko, Antoine Caubrière, Benjamin Lecouteux, Yannick Estève, and Laurent Besacier. 2020b. ON-TRAC consortium for end-to-end and simultaneous speech translation challenge tasks at IWSLT 2020. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 35–43, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.iwslt-1.2

Claudio Fantinuoli and Bianca Prandi. 2021. Towards the evaluation of automatic simultaneous speech translation from a communicative perspective. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 245–254, Bangkok, Thailand (online). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.iwslt-1.29

Claudio Fantinuoli and Xiaoman Wang. 2024. Exploring the correlation between human and machine evaluation of simultaneous speech translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 327–336, Ses327–336, Sheffield, UK. European Association for Machine Translation (EAMT).

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. https://doi.org/10.18653/v1/2021.acl-long.505

Luciana Ferrer, Elizabeth Shriberg, and Andreas Stolcke. 2003. A prosody-based approach to end-of-utterance detection that does not require speech recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 (ICASSP'03)*. IEEE. https://doi.org/10.1109/ICASSP.2003.1198854

Biao Fu, Minpeng Liao, Kai Fan, Zhongqiang Huang, Boxing Chen, Yidong Chen, and Xiaodong Shi. 2023. Adapting offline speech translation models for streaming with future-aware distillation and inference. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16600–16619, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.1033

Christian Fügen, Muntsin Kolss, Dietmar Bernreuther, Matthias Paulik, Sebastian Stuker, Stephan Vogel, and Alex Waibel. 2006a. Open domain speech recognition & translation: Lectures and speeches. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. https://doi.org/10.1109/ICASSP.2006.1660084

Christian Fügen, Muntsin Kolss, Matthias Paulik, and Alex Waibel. 2006b. Open domain speech translation: From seminars and speeches to lectures. In *TC-STAR Workshop on Speech to Speech Translation, Barcelona, Spain*, pages 81–86.

Christian Fügen, Alex Waibel, and Muntsin Kolss. 2007. Simultaneous translation of lectures and speeches. *Machine Translation*, 21:209–252. https://doi.org/10.1007/s10590-008-9047-0

Christian Fügen. 2009. *A System for Simultaneous Translation of Lectures and Speeches*. Ph.D. thesis, Universität Karlsruhe (TH). https://doi.org/10.5445/IR/1000013594

Tomoki Fujita, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2013. Simple, lexicalized choice of translation timing for simultaneous speech translation. In *Proceedings of Interspeech 2013*, pages 3487–3491. https://doi.org/10.21437/Interspeech.2013-615

Ryo Fukuda, Yuka Ko, Yasumasa Kano, Kosuke Doi, Hirotaka Tokuyama, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. 2022a. NAIST simultaneous speech-to-text translation system for IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 286–292, Dublin, Ireland (in-person and online). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.iwslt-1.25

Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Yuka Ko, Tomoya Yanagita, Kosuke Doi, Mana Makinae, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. 2023. NAIST simultaneous speech-to-speech translation system for IWSLT 2023. In *Poceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 330–340, Toronto, Canada (in-person and online). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.iwslt-1.31

Ryo Fukuda, Katsuhito Sudoh, and Satoshi Nakamura. 2022b. Speech segmentation optimization using segmented bilingual speech corpus for end-to-end speech translation. In *Proceedings of Interspeech 2022*, pages 121–125. https://doi.org/10.21437/Interspeech.2022-11382

Marco Gaido, Mattia A. Di Gangi, Matteo Negri, Mauro Cettolo, and Marco Turchi. 2020. Contextualized translation of automatically segmented speech. In *Proceedings of Interspeech 2020*, pages 1471–1475. https://doi.org/10.21437/Interspeech.2020-2860

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, Trento, Italy. Association for Computational Linguistics.

Marco Gaido, Sara Papi, Dennis Fucci, Giuseppe Fiameni, Matteo Negri, and Marco Turchi. 2022. Efficient yet competitive speech translation: FBK@IWSLT2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 177–189, Dublin, Ireland (in-person and online). Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.iwslt-1.13`

Marco Gaido, Sara Papi, Matteo Negri, and Luisa Bentivogli. 2024. Speech translation with speech foundation models and large language models: What is there and what is missing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14760–14778, Bangkok, Thailand. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2024.acl-long.789`

Marco Gaido, Sara Papi, Matteo Negri, and Marco Turchi. 2023. Joint speech translation and named entity recognition. In *Proceedings of INTERSPEECH 2023*, pages 47–51. `https://doi.org/10.21437/Interspeech.2023-1767`

Karl R. Gegenfurtner. 2016. The interaction between vision and eye movements. *Perception*, 45(12):1333–1357. `https://doi.org/10.1177/0301006616657097`, PubMed: 27383394

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*, ICML '06, pages 369–376, New York, NY, USA. Association for Computing Machinery. `https://doi.org/10.1145/1143844.1143891`

Alvin Grissom II, He He, Jordan Boyd-Graber, John Morgan, and Hal Daumé III. 2014. Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1342–1352, Doha, Qatar. Association for Computational Linguistics. `https://doi.org/10.3115/v1/D14-1140`

Bao Guo, Mengge Liu, Wen Zhang, Hexuan Chen, Chang Mu, Xiang Li, Jianwei Cui, Bin Wang, and Yuhang Guo. 2022. The xiaomi text-to-text simultaneous speech translation system for IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 216–224, Dublin, Ireland (in-person and online). Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.iwslt-1.17`

Jiaxin Guo, Daimeng Wei, Zhanglin Wu, Zongyao Li, Zhiqiang Rao, Minghan Wang, Hengchao Shang, Xiaoyu Chen, Zhengzhe Yu, Shaojun Li, Yuhao Xie, Lizhi Lei, and Hao Yang. 2023. The HW-TSC's simultaneous speech-to-text translation system for IWSLT 2023 evaluation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 376–382, Toronto, Canada (in-person and online). Association for Computational Linguistics. `https://doi.org/10.18653/v1/2023.iwslt-1.35`

Jiaxin Guo, Zhanglin Wu, Zongyao Li, Hengchao Shang, Daimeng Wei, Xiaoyu Chen, Zhiqiang Rao, Shaojun Li, and Hao Yang. 2024. R-bi: Regularized batched inputs enhance incremental decoding framework for low-latency simultaneous speech translation. *arXiv preprint arXiv:2401.05700*.

Hou Jeung Han, Mohd Abbas Zaidi, Sathish Reddy Indurthi, Nikhil Kumar Lakumarapu, Beomseok Lee, and Sangha Kim. 2020. End-to-end simultaneous translation system for IWSLT2020 using modality agnostic meta-learning. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 62–68, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.iwslt-1.5`

Wuwei Huang, Mengge Liu, Xiang Li, Yanzhi Tian, Fengyu Yang, Wen Zhang, Jian Luan, Bin Wang, Yuhang Guo, and Jinsong Su. 2023. The xiaomi AI lab's speech translation systems for IWSLT 2023 offline task, simultaneous task and speech-to-speech task. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 411–419, Toronto, Canada (in-person and online). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.iwslt-1.39

W. Ronny Huang, Shuo-Yiin Chang, David Rybach, Tara Sainath, Rohit Prabhavalkar, Cal Peyser, Zhiyun Lu, and Cyril Allauzen. 2022. E2e segmenter: Joint segmenting and decoding for long-form asr. In *Interspeech 2022*, pages 4995–4999. https://doi.org/10.21437/Interspeech.2022-38

Christian Huber, Tu Anh Dinh, Carlos Mullov, Ngoc-Quan Pham, Thai Binh Nguyen, Fabian Retkowski, Stefan Constantin, Enes Ugan, Danni Liu, Zhaolin Li, Sai Koneru, Jan Niehues, and Alexander Waibel. 2023. End-to-end evaluation for low-latency simultaneous speech translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 12–20, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-demo.2

Christian Huber, Enes Yavuz Ugan, and Alexander Waibel. 2022. Code-switching without switching: Language agnostic end-to-end speech translation. *arXiv preprint arXiv:2210.01512*.

Min-Jae Hwang, Ilia Kulikov, Benjamin Peloquin, Hongyu Gong, Peng-Jen Chen, and Ann Lee. 2024. Textless acoustic model with self-supervised distillation for noise-robust expressive speech-to-speech translation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15524–15541. Bangkok, Thailand and virtual meeting. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.917

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. https://doi.org/10.18653/v1/2021.iwslt-1.10

Sathish Reddy Indurthi, Mohd Abbas Zaidi, Beomseok Lee, Nikhil Kumar Lakumarapu, and Sangha Kim. 2022. Language model augmented monotonic attention for simultaneous translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 38–45, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.3

Javier Iranzo-Sánchez, Adrià Giménez Pastor, Joan Albert Silvestre-Cerdà, Pau Baquero-Arnal, Jorge Civera Saiz, and Alfons Juan. 2020. 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. https://doi.org/10.18653/v1/2020.emnlp-main.206

Javier Iranzo-Sánchez, Javier Jorge Cano, Alejandro Pérez-González-de Martos, Adrián Giménez Pastor, Gonçal Garcés Díaz-Munío, Pau Baquero-Arnal, Joan Albert Silvestre-Cerdà, Jorge Civera Saiz, Albert Sanchis, and Alfons Juan. 2022. MLLP-VRAIN UPV systems for the IWSLT 2022 simultaneous speech translation and speech-to-speech translation tasks. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 255–264, Dublin, Ireland (in-person and online). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.iwslt-1.22

Javier Iranzo-Sánchez, Jorge Iranzo-Sánchez, Adrià Giménez, Jorge Civera, and Alfons Juan. 2024. Segmentation-free streaming machine translation. *Transactions of the Association for Computational Linguistics*, 12:1104–1121. https://doi.org/10.1162/tacl_a_00691

Javier Iranzo-Sánchez, Javier Jorge, Pau Baquero-Arnal, Joan Albert Silvestre-Cerdà, Adrià Giménez, Jorge Civera, Albert Sanchis, and Alfons Juan. 2021. Streaming cascade-based speech translation leveraged by a direct segmentation model. *Neural Networks*, 142:303–315. https://doi.org/10.1016/j.neunet.2021.05.013, PubMed: 34082286

Christopher Irvin. 2021. Student insights related to the use of simultaneous speech translation for video lectures in a university english course. *STEM Journal*. https://doi.org/10.16875/stem.2021.22.4.59

Dávid Javorský, Dominik Macháček, and Ondřej Bojar. 2022. Continuous rating as reliable human evaluation of simultaneous speech translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 154–164, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. 2023. Average token delay: A latency metric for simultaneous translation. In *Proceedings of INTERSPEECH 2023*, pages 4469–4473. https://doi.org/10.21437/Interspeech.2023-933

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020. Is 42 the answer to everything in subtitling-oriented speech translation? In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 209–219, Online. https://doi.org/10.18653/v1/2020.iwslt-1.26

Alina Karakanta, Sara Papi, Matteo Negri, and Marco Turchi. 2021. Simultaneous speech translation for live subtitling: From delay to display. In *Proceedings of the 1st Workshop on Automatic Spoken Language Translation in Real-World Settings (ASLTRW)*, pages 35–48, Virtual. Association for Machine Translation in the Americas.

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)*, pages 24–34, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-6503

Yuka Ko, Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. 2023. Tagged end-to-end simultaneous speech translation training using simultaneous interpretation data. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 363–375, Toronto, Canada (in-person and online). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.iwslt-1.34

Yuka Ko, Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Tomoya Yanagita, Kosuke Doi, Mana Makinae, Haotian Tan, Makoto Sakai, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. 2024. NAIST simultaneous speech translation system for IWSLT 2024. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 170–182, Bangkok, Thailand (in-person and online). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.iwslt-1.23

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.110

Muntsin Kolss, Matthias Wölfel, Florian Kraft, Jan Niehues, Matthias Paulik, and Alex Waibel. 2008. Simultaneous German-English lecture translation. In *Proceedings of the 5th International Workshop on Spoken Language Translation: Papers*, pages 174–181, Waikiki, Hawaii.

Phillip A. Laplante. 1992. *Real-time Systems Design and Analysis: An Engineer's Handbook*. IEEE Press.

Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Heriberto Cuayáhuitl, and Björn W. Schuller. 2023. Sparks of large audio models: A survey and outlook. *arXiv preprint arXiv:2308.12792*.

Zecheng Li, Yue Sun, and Haoze Li. 2022. System description on automatic simultaneous translation workshop. In *Proceedings of the Third Workshop on Automatic Simultaneous Translation*, pages 18–21, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.autosimtrans-1.3

Dan Liu, Mengge Du, Xiaoxi Li, Yuchen Hu, and Lirong Dai. 2021a. The USTC-NELSLIP systems for simultaneous speech translation task at IWSLT 2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 30–38, Bangkok, Thailand (online). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.iwslt-1.2

Dan Liu, Mengge Du, Xiaoxi Li, Ya Li, and Enhong Chen. 2021b. 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. https://doi.org/10.18653/v1/2021.emnlp-main.4

Xiaoqian Liu, Guoqiang Hu, Yangfan Du, Erfeng He, YingFeng Luo, Chen Xu, Tong Xiao, and Jingbo Zhu. 2024. Recent advances in end-to-end simultaneous speech translation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8142–8150, International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2024/900

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, Cambridge, MA. Association for Computational Linguistics.

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*.

Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *International Conference on Learning Representations*.

Qianxi Lv and Junying Liang. 2019. Is consecutive interpreting easier than simultaneous interpreting? – A corpus-based study of lexical simplification in interpretation. *Perspectives*, 27(1):91–106. https://doi.org/10.1080/0907676X.2018.1498531

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. https://doi.org/10.18653/v1/P19-1289

Xutai Ma, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. 2020a. 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. https://doi.org/10.18653/v1/2020.emnlp-demos.19

Xutai Ma, Juan Pino, and Philipp Koehn. 2020b. SimulMT to SimulST: Adapting simultaneous text translation to end-to-end simultaneous speech translation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 582–587, Suzhou, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.aacl-main.58

Xutai Ma, Anna Sun, Siqi Ouyang, Hirofumi Inaguma, and Paden Tomasello. 2023. Efficient monotonic multihead attention. *arXiv preprint arXiv:2312.04515*.

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. https://doi.org/10.1109/ICASSP39728.2021.9414897

Zhengrui Ma, Qingkai Fang, Shaolei Zhang, Shoutao Guo, Yang Feng, and Min Zhang.

2024. A non-autoregressive generation framework for end-to-end simultaneous speech-to-any translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1557–1575, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.85

Dominik Macháček, Ondřej Bojar, and Raj Dabre. 2023. 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. https://doi.org/10.18653/v1/2023.iwslt-1.12

Dominik Macháček, Jonáš Kratochvíl, Sangeet Sagar, Matúš Žilinec, Ondřej Bojar, Thai-Son Nguyen, Felix Schneider, Philip Williams, and Yuekun Yao. 2020. ELITR non-native speech translation at IWSLT 2020. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 200–208, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.iwslt-1.25

Dominik Machácek and Ondrej Bojar. 2020. Presenting simultaneous translation in limited space. In *Proceedings of the 20th Conference Information Technologies - Applications and Theory (ITAT 2020)*, pages 32–37, Košice, Slovakia. Tomáš Horváth.

Evgeny Matusov, Stephan Kanthak, and Hermann Ney. 2006. Integrating speech recognition and machine translation: Where do we stand? In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V. https://doi.org/10.1109/ICASSP.2006.1661501

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.

Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou. 2019. Customizing neural machine translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy. https://doi.org/10.18653/v1/W19-5209

Niko Moritz, Takaaki Hori, and Jonathan Le. 2020. Streaming automatic speech recognition with the transformer model. In *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6074–6078. https://doi.org/10.1109/ICASSP40776.2020.9054476

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. https://doi.org/10.18653/v1/N16-3017

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. https://doi.org/10.1109/ASRU46091.2019.9003913

Ha Nguyen, Yannick Estève, and Laurent Besacier. 2021a. An empirical study of end-to-end simultaneous speech translation decoding strategies. In *ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7528–7532. https://doi.org/10.1109/ICASSP39728.2021.9414276

Ha Nguyen, Yannick Estève, and Laurent Besacier. 2021b. Impact of encoding and segmentation strategies on end-to-end simultaneous speech translation. In *Interspeech 2021*, pages 2371–2375. https://doi.org/10.21437/Interspeech.2021-608

Jan Niehues, Rolando Cattoni, Sebastian Stüker, Mauro Cettolo, Marco Turchi, and Marcello

Federico. 2018a. The IWSLT 2018 evaluation campaign. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 2–6, Brussels. International Conference on Spoken Language Translation.

Jan Niehues, Rolando Cattoni, Sebastian Stüker, Matteo Negri, Marco Turchi, Thanh-Le Ha, Elizabeth Salesky, Ramon Sanabria, Loic Barrault, Lucia Specia, and Marcello Federico. 2019. The IWSLT 2019 evaluation campaign. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics. https://doi.org/10.21437/Interspeech.2016-154

Jan Niehues, Thai Son Nguyen, Eunah Cho, Thanh-Le Ha, Kevin Kilgour, Markus Müller, Matthias Sperber, Sebastian Stüker, and Alex Waibel. 2016. Dynamic Transcription for Low-Latency Speech Translation. In *Proceedings of Interspeech 2016*, pages 2513–2517. https://doi.org/10.21437/Interspeech.2018-1055

Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber, and Alex Waibel. 2018b. Low-latency neural speech translation. In *Interspeech 2018*, pages 1293–1297.

Sashi Novitasari, Takashi Fukuda, and Gakuto Kurata. 2022. Improving asr robustness in noisy condition through vad integration. In *Interspeech 2022*, pages 3784–3788. https://doi.org/10.21437/Interspeech.2022-260

Sashi Novitasari, Sakriani Sakti, and Satoshi Nakamura. 2021. Neural incremental speech recognition toward real-time machine speech translation. *IEICE Transactions on Information and Systems*, E104.D(12):2195–2208. https://doi.org/10.1587/transinf.2021EDP7014

Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Optimizing segmentation strategies for simultaneous speech translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 551–556, Baltimore, Maryland. Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-2090

Motoi Omachi, Brian Yan, Siddharth Dalmia, Yuya Fujita, and Shinji Watanabe. 2023. Align, write, re-order: Explainable end-to-end speech translation via operation sequence generation. In *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095896

Sara Papi, Marco Gaido, and Matteo Negri. 2023a. Direct models for simultaneous translation and automatic subtitling: FBK@IWSLT2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 159–168, Toronto, Canada (in-person and online). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.iwslt-1.11

Sara Papi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2024a. SimulSeamless: FBK at IWSLT 2024 simultaneous speech translation. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 72–79, Bangkok, Thailand (in-person and online). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.iwslt-1.11

Sara Papi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2024b. StreamAtt: Direct streaming speech-to-text translation with attention-based audio history selection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3692–3707, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.202

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2022a. Does simultaneous speech translation need simultaneous models? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 141–153, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.11

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2022b. 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. https://doi.org/10.18653/v1/2022.autosimtrans-1.2

Sara Papi, Alina Karakanta, Matteo Negri, and Marco Turchi. 2022c. Dodging the data bottleneck: Automatic subtitling with automatically segmented ST corpora. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 480–487, Online only. https://doi.org/10.18653/v1/2022.aacl-short.59

Sara Papi, Matteo Negri, and Marco Turchi. 2021. Visualization: The missing factor in simultaneous speech translation. In *CEUR Workshop Proceedings*, volume 3033.

Sara Papi, Matteo Negri, and Marco Turchi. 2023b. Attention as a guide for simultaneous speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13340–13356, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.745

Sara Papi, Marco Turchi, and Matteo Negri. 2023c. AlignAtt: Using attention-based audio-translation alignments as a guide for simultaneous speech translation. In *Proceedings of INTERSPEECH 2023*, pages 3974–3978. https://doi.org/10.21437/Interspeech.2023-170

Sara Papi, Peidong Wang, Junkun Chen, Jian Xue, Naoyuki Kanda, Jinyu Li, and Yashesh Gaur. 2024c. Leveraging timestamp information for serialized joint streaming recognition and translation. In *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10381–10385. https://doi.org/10.1109/ICASSP48485.2024.10447565

Sara Papi, Peidong Wang, Junkun Chen, Jian Xue, Jinyu Li, and Yashesh Gaur. 2023d. Token-level serialized output training for joint streaming asr and st leveraging textual alignments. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. https://doi.org/10.1109/ASRU57964.2023.10389715

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth Narayanan. 2022. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317. https://doi.org/10.1016/j.csl.2021.101317

Matthias Paulik and Alex Waibel. 2010. Rapid development of speech translation using consecutive interpretation. In *Proceedings of Interspeech 2010*, pages 2534–2537. https://doi.org/10.21437/Interspeech.2010-680

Elisa Perego, Fabio Del Missier, Marco Porta, and Mauro Mosconi. 2010. The cognitive effectiveness of subtitle processing. *Media Psychology*, 13(3):243–272. https://doi.org/10.1080/15213269.2010.502873

Peter Polák. 2023. Long-form simultaneous speech translation: Thesis proposal. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 64–74, Nusa Dua, Bali. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.ijcnlp-srw.9

Peter Polák and Ondřej Bojar. 2023. Long-form end-to-end speech translation via latent alignment segmentation. *arXiv preprint arXiv:2309.11384*. https://doi.org/10.1109/SLT61566.2024.10832264

Peter Polák, Danni Liu, Ngoc-Quan Pham, Jan Niehues, Alexander Waibel, and Ondřej Bojar. 2023. 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. https://doi.org/10.18653/v1/2023.iwslt-1.37

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. https://doi.org/10.18653/v1/2022.iwslt-1.24

Peter Polák, Brian Yan, Shinji Watanabe, Alex Waibel, and Ondrej Bojar. 2023. Incremental blockwise beam search for simultaneous speech translation with controllable quality-latency tradeoff. In *Proceedings of INTERSPEECH 2023*, pages 3979–3983. https://doi.org/10.21437/Interspeech.2023-2225

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. https://doi.org/10.18653/v1/2020.iwslt-1.9

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.

Matthew Raffel and Lizhong Chen. 2023. Implicit memory transformer for computationally efficient simultaneous speech translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12900–12907, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.816

Matthew Raffel, Drew Penney, and Lizhong Chen. 2023. Shiftable context: Addressing training-inference context mismatch in simultaneous speech translation. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Dhevi J. Rajendran, Andrew T. Duchowski, Pilar Orero, Juan Martínez, and Pablo Romero-Fresco. 2013. Effects of text chunking on subtitling: A quantitative and qualitative examination. *Perspectives*, 21(1):5–21. https://doi.org/10.1080/0907676X.2012.722651

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, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Yi Ren, Jinglin Liu, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. SimulSpeech: End-to-end simultaneous speech to text translation. In *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics, pages 3787–3796, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.350

Pablo Romero-Fresco. 2010. Standing on quicksand: Hearing viewers' comprehension and reading patterns of respoken subtitles for the news. In *New insights into audiovisual translation and media accessibility*, pages 175–194. Brill. https://doi.org/10.1163/9789042031814_014

Pablo Romero-Fresco. 2011. *Subtitling through Speech Recognition: Respeaking*. Routledge.

Koichiro Ryu, Shigeki Matsubara, and Yasuyoshi Inagaki. 2006. Simultaneous English-Japanese spoken language translation based on incremental dependency parsing and transfer. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 683–690, Sydney, Australia. Association for Computational Linguistics. https://doi.org/10.3115/1273073.1273161

Akiko Sakamoto, Kazuhiko Abe, Kazuo Sumita, and Satoshi Kamatani. 2013. Evaluation of a simultaneous interpretation system and analysis of speech log for user experience assessment. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*, Heidelberg, Germany.

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. https://doi.org/10.18653/v1/2020.iwslt-1.28

Sukanta Sen, Ondřej Bojar, and Barry Haddow. 2022. Simultaneous translation for unsegmented input: A sliding window approach. *arXiv preprint arXiv:2210.09754*.

Hassan Shavarani, Maryam Siahbani, Ramtin Mehdizadeh Seraj, and Anoop Sarkar. 2015. Learning segmentations that balance latency versus quality in spoken language translation. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Papers*, pages 217–224, Da Nang, Vietnam.

Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2013. Constructing a speech translation system using simultaneous interpretation data. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*, Heidelberg, Germany.

Maryam Siahbani, Hassan Shavarani, Ashkan Alinejad, and Anoop Sarkar. 2018. Simultaneous translation using optimized segmentation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 154–167, Boston, MA. Association for Machine Translation in the Americas.

Mark Sinclair, Peter Bell, Alexandra Birch, and Fergus McInnes. 2014. A semi-Markov model for speech segmentation with an utterance-break prior. In *Proceedings of Interspeech 2014*, pages 2351–2355. https://doi.org/10.21437/Interspeech.2014-511

Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. 1999. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3. https://doi.org/10.1109/97.736233

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. https://doi.org/10.18653/v1/2020.acl-main.661

Fred W. M. Stentiford and Martin G. Steer. 1988. Machine translation of speech. *British Telecom Technology Journal*, 6(2):116–122.

Shashank Subramanya and Jan Niehues. 2022. Multilingual simultaneous speech translation. *arXiv preprint arXiv:2203.14835*.

Weiting Tan, Yunmo Chen, Tongfei Chen, Guanghui Qin, Haoran Xu, Heidi C. Zhang, Benjamin Van Durme, and Philipp Koehn. 2024. Streaming sequence transduction through dynamic compression. *arXiv preprint arXiv:2402.01172*.

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. https://doi.org/10.18653/v1/2023.acl-long.695

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6). https://doi.org/10.1145/3530811

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. https://doi.org/10.18653/v1/W17-4811

Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta R. Costa-jussà. 2022. SHAS: Approaching optimal segmentation for end-to-end speech translation. In *Proceedings of Interspeech 2022*, pages 106–110. https://doi.org/10.21437/Interspeech.2022-59

Alexander H. Waibel. 2004. Speech translation: Past, present and future. In *Interspeech*. https://doi.org/10.21437/Interspeech.2004-156

Alexander H. Waibel, Ajay N. Jain, Arthur E. McNair, Hiroaki Saito, Alexander Hauptmann, and Joe Tebelskis. 1991. Janus: A speech-to-speech translation system using connectionist and symbolic processing strategies. *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, 793–796 vol. 2. https://doi.org/10.1109/ICASSP.1991.150456

Jinhan Wang, Xiaosu Tong, Jinxi Guo, Di He, and Roland Maas. 2022a. Vadoi: Voice-activity-detection overlapping inference for end-to-end long-form speech recognition. In *ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6977–6981. https://doi.org/10.1109/ICASSP43922.2022.9746873

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. 2022b. 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. https://doi.org/10.18653/v1/2022.iwslt-1.21

Peidong Wang, Eric Sun, Jian Xue, Yu Wu, Long Zhou, Yashesh Gaur, Shujie Liu, and Jinyu Li. 2023. LAMASSU: A streaming language-agnostic multilingual speech recognition and translation model using neural transducers. In *Proceedings of INTERSPEECH 2023*, pages 57–61. https://doi.org/10.21437/Interspeech.2023-2004

Xiaolin Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2016. An efficient and effective online sentence segmenter for simultaneous interpretation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 139–148, Osaka, Japan. The COLING 2016 Organizing Committee.

Xiaolin Wang, Masao Utiyama, and Eiichiro Sumita. 2019. Online sentence segmentation for simultaneous interpretation using multi-shifted recurrent neural network. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 1–11, Dublin, Ireland. European Association for Machine Translation.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. In *Interspeech 2017*, pages 2625–2629. https://doi.org/10.21437/Interspeech.2017-503

Orion Weller, Matthias Sperber, Christian Gollan, and Joris Kluivers. 2021. Streaming models for joint speech recognition and translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2533–2539, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.216

Orion Weller, Matthias Sperber, Telmo Pires, Hendra Setiawan, Christian Gollan, Dominic Telaar, and Matthias Paulik. 2022. End-to-end speech translation for code switched speech. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1435–1448, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.113

Patrick Wilken, Tamer Alkhouli, Evgeny Matusov, and Pavel Golik. 2020. Neural simultaneous speech translation using alignment-based chunking. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 237–246, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.iwslt-1.29

Matthias Wolfel, Muntsin Kolss, Florian Kraft, Jan Niehues, Matthias Paulik, and Alex Waibel. 2008. Simultaneous machine translation of german lectures into english: Investigating research challenges for the future. In *2008 IEEE Spoken Language Technology Workshop*, pages 233–236. https://doi.org/10.1109/SLT.2008.4777883

Krzysztof Wołk and Krzysztof Marasek. 2014. Real-time statistical speech translation. In *New Perspectives in Information Systems and Technologies, Volume 1*, pages 107–113. Springer. https://doi.org/10.1007/978-3-319-05951-8_11

Monika Woszczyna, Matthew Broadhead, Donna Gates, Marsal Gavalda, Alon Lavie, Lori Levin, and Alex Waibel. 1998. A modular approach to spoken language translation for large domains. In *Machine Translation and the Information Soup: Third Conference of the Association for Machine Translation in the Americas AMTA'98 Langhorne, PA, USA, October 28–31, 1998 Proceedings 3*, pages 31–40. Springer. https://doi.org/10.1007/3-540-49478-2_3

Chunyang Wu, Yongqiang Wang, Yangyang Shi, Ching-Feng Yeh, and Frank Zhang. 2020. Streaming transformer-based acoustic models using self-attention with augmented memory. In *Proceedings of Interspeech 2020*, pages 2132–2136. https://doi.org/10.21437/Interspeech.2020-2079

Hao Xiong, Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Dutongchuan: Context-aware translation model for simultaneous interpreting. *arXiv preprint arXiv:1907.12984*.

Jian Xue, Peidong Wang, Jinyu Li, Matt Post, and Yashesh Gaur. 2022. Large-scale streaming end-to-end speech translation with neural transducers. In *Proceedings of Interspeech 2022*, pages 3263–3267. https://doi.org/10.21437/Interspeech.2022-10953

Jian Xue, Peidong Wang, Jinyu Li, and Eric Sun. 2023. A weakly-supervised streaming multilingual speech model with truly zero-shot capability. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. https://doi.org/10.1109/ASRU57964.2023.10389799

Brian Yan, Jiatong Shi, Soumi Maiti, William Chen, Xinjian Li, Yifan Peng, Siddhant Arora, and Shinji Watanabe. 2023. CMU's IWSLT 2023 simultaneous speech translation system. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 235–240, Toronto, Canada (in-person and online). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.iwslt-1.20

Mu Yang, Naoyuki Kanda, Xiaofei Wang, Junkun Chen, Peidong Wang, Jian Xue, Jinyu Li, and Takuya Yoshioka. 2024. Diarist: Streaming speech translation with speaker diarization. In *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10866–10870. https://doi.org/10.1109/ICASSP48485.2024.10446050

Yuekun Yao and Barry Haddow. 2020. Dynamic masking for improved stability in online spoken language translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 123–136, Virtual. Association for Machine Translation in the Americas.

Mahsa Yarmohammadi, Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Baskaran Sankaran. 2013. Incremental segmentation and decoding strategies for simultaneous translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1032–1036, Nagoya, Japan. Asian Federation of Natural Language Processing.

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. https://doi.org/10.1109/ICASSP40776.2020.9054358

Mohd Abbas Zaidi, Beomseok Lee, Sangha Kim, and Chanwoo Kim. 2021. Decision attentive regularization to improve simultaneous speech translation systems. *arXiv preprint arXiv:2110.15729*. https://doi.org/10.21437/Interspeech.2022-10617

Mohd Abbas Zaidi, Beomseok Lee, Sangha Kim, and Chanwoo Kim. 2022. Cross-modal decision regularization for simultaneous speech translation. In *Proceedings of Interspeech 2022*, pages 116–120. https://doi.org/10.21437/Interspeech.2022-10617

Xingshan Zeng, Liangyou Li, and Qun Liu. 2021. RealTranS: End-to-end simultaneous speech translation with convolutional weighted-shrinking transformer. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2461–2474, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.218

Xingshan Zeng, Pengfei Li, Liangyou Li, and Qun Liu. 2022. End-to-end simultaneous speech translation with pretraining and distillation: Huawei Noah's system for AutoSimTranS 2022. In *Proceedings of the Third Workshop on Automatic Simultaneous Translation*, pages 25–33, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.autosimtrans-1.5

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, Online. https://doi.org/10.18653/v1/2021.acl-long.200

Linlin Zhang, Kai Fan, Jiajun Bu, and Zhongqiang Huang. 2023a. Training simultaneous speech translation with robust and random wait-k-tokens strategy. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7814–7831, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.484

Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar. 2020. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss. In *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7829–7833. https://doi.org/10.1109/ICASSP40776.2020.9053896

Ruiqing Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2022. Learning adaptive segmentation policy for end-to-end simultaneous translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7862–7874, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.542

Shaolei Zhang, Qingkai Fang, Shoutao Guo, Zhengrui Ma, Min Zhang, and Yang Feng. 2024. StreamSpeech: Simultaneous speech-to-speech translation with multi-task learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8964–8986, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.485

Shaolei Zhang and Yang Feng. 2022. Information-transport-based policy for simultaneous translation. In *Proceedings of the 2022*

Conference on Empirical Methods in Natural Language Processing, pages 992–1013, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.65

Shaolei Zhang and Yang Feng. 2023. End-to-end simultaneous speech translation with differentiable segmentation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7659–7680, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.485

Shaolei Zhang and Yang Feng. 2024. Unified segment-to-segment framework for simultaneous sequence generation. *Advances in Neural Information Processing Systems*, 36.

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

Renjie Zheng, Mingbo Ma, Baigong Zheng, Kaibo Liu, and Liang Huang. 2020. Opportunistic decoding with timely correction for simultaneous translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 437–442, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.42

Qinpei Zhu, Renshou Wu, Guangfeng Liu, Xinyu Zhu, Xingyu Chen, Yang Zhou, Qingliang Miao, Rui Wang, and Kai Yu. 2022. The AISP-SJTU simultaneous translation system for IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 208–215, Dublin, Ireland (in-person and online). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.iwslt-1.16

A Categorized Papers

The papers retrieved for the statistics provided in §4 are obtained by searching on Semantic Scholar using the following queries:[11]

[11] Accessed July 6, 2024.

Query	**#papers**
`simultaneous+speech+translation`	265
`streaming+speech+translation`	218
`real-time+speech+translation`	265
`online+speech+translation`	250
`simultaneous+spoken+language+translation`	181
`streaming+spoken+language+translation`	85
`real-time+spoken+language+translation`	218
`online+spoken+language+translation`	69

Table 2: Queries used for research on the Semantic Scholar database with their corresponding number of resulting papers.

Notice that querying for ``speech'' already includes the results for ``speech-to-text'' and similar combinations. Moreover, since we are interested in trends in SimulST systems, we include only papers proposing models (i.e., excluding corpora, surveys, and metrics) and providing results for the speech-to-text task (i.e., speech-to-speech and/or text-to-text are not considered). Only papers written in English and with an open-access version have been considered.

The analysis resulted in 110 papers, categorized following our taxonomy (Figure 2) and reported in the following in chronological order. Notice that, in some cases, the number of papers on the various dichotomies does not sum to 110 since some work proposes, for instance, both cascade and direct models and appear in both categories.

A.1 By Input Type

A.1.1 Bounded Speech (90 papers)

Automatic Pre-Segmentation (2 papers). Kolss et al. (2008), Shimizu et al. (2013)

Gold Pre-Segmentation (88 papers). Ryu et al. (2006), Kolss et al. (2008), Fujita et al. (2013), Rangarajan Sridhar et al. (2013), Yarmohammadi et al. (2013), Oda et al. (2014), Wołk and Marasek (2014), Cho et al. (2015), Shavarani et al. (2015), Cho et al. (2017), Siahbani et al. (2018), Xiong et al. (2019), Arivazhagan et al. (2020a), Bahar et al. (2020), Elbayad et al. (2020a), Elbayad et al. (2020b), Han et al. (2020), Ma et al. (2020b), Ren et al. (2020), Wilken et al. (2020), Yao and Haddow (2020), Nguyen et al. (2021a), Ma et al. (2021),[12] Bahar et al. (2021), Chen et al. (2021), Karakanta et al. (2021), Liu et al. (2021b), Liu et al. (2021a), Nguyen et al. (2021b), Novitasari et al. (2021), Weller et al. (2021), Zaidi et al. (2021), Zeng et al. (2021), Chang and yi Lee (2022), Deng et al. (2022), Dong et al. (2022), Fukuda et al. (2022a), Gaido et al. (2022), Guo et al. (2022), Indurthi et al. (2022), Iranzo-Sánchez et al. (2022), Li et al. (2022), Papi et al. (2022a), Polák et al. (2022), Subramanya and Niehues (2022), Wang et al. (2022b), Xue et al. (2022), Zaidi et al. (2022), Zeng et al. (2022), Zhang et al. (2022), Zhang and Feng (2022), Zhu et al. (2022), Omachi et al. (2023), Chen et al. (2023), Xue et al. (2023), Raffel et al. (2023), Alastruey et al. (2023), Barrault et al. (2023), Fu et al. (2023), Fukuda et al. (2023), Gaido et al. (2023), Guo et al. (2023), Huang et al. (2023), Ko et al. (2023), Ma et al. (2023), Papi et al. (2023d), Papi et al. (2023c), Papi et al. (2023b), Papi et al. (2023a), Polák et al. (2023), Polák et al. (2023), Raffel and Chen (2023), Tang et al. (2023), Wang et al. (2023), Yan et al. (2023), Zhang et al. (2023a), Zhang and Feng (2023), Yang et al. (2024), Chen et al. (2024), Deng and Woodland (2024), Guo et al. (2024), Ko et al. (2024), Ma et al. (2024), Papi et al. (2024c), Papi et al. (2024a), Tan et al. (2024), Zhang et al. (2024), Zhang and Feng (2024)

A.1.2 Unbounded Speech (20 papers)

Simultaneous (Automatic) Segmentation (14 papers). Fügen et al. (2006a), Fügen et al. (2007), Wolfel et al. (2008), Fügen (2009), Cho et al. (2013), Müller et al. (2016), Niehues et al. (2016), Wang et al. (2016), Wang et al. (2019), Arivazhagan et al. (2020b), Iranzo-Sánchez et al. (2020), Macháček et al. (2020), Bojar et al. (2021), Iranzo-Sánchez et al. (2021),

Segmentation-free (6 papers). Schneider and Waibel (2020), Amrhein and Haddow (2022), Sen et al. (2022), Iranzo-Sánchez et al. (2024), Polák (2023), Papi et al. (2024b)

A.1.3 Undefined (1 paper)

Dessloch et al. (2018)

[12]Unbounded speech theoretically possible but not tested.

A.2 By Architecture

A.2.1 Direct (64 papers)

Han et al. (2020), Ma et al. (2020b), Ren et al. (2020), Nguyen et al. (2021a), Ma et al. (2021), Chen et al. (2021), Karakanta et al. (2021), Liu et al. (2021b), Liu et al. (2021a), Nguyen et al. (2021b), Zaidi et al. (2021), Zeng et al. (2021), Amrhein and Haddow (2022), Chang and yi Lee (2022), Deng et al. (2022), Dong et al. (2022), Fukuda et al. (2022a), Gaido et al. (2022), Papi et al. (2022a), Polák et al. (2022), Subramanya and Niehues (2022), Wang et al. (2022b), Xue et al. (2022), Zaidi et al. (2022), Zhang et al. (2022), Zhang and Feng (2022), Zhu et al. (2022), Omachi et al. (2023), Chen et al. (2023), Xue et al. (2023), Raffel et al. (2023), Alastruey et al. (2023), Barrault et al. (2023), Fu et al. (2023), Fukuda et al. (2023), Gaido et al. (2023), Huang et al. (2023), Ko et al. (2023), Ma et al. (2023), Papi et al. (2023d), Papi et al. (2023c), Papi et al. (2023b), Papi et al. (2023a), Polák (2023), Polák et al. (2023), Polák et al. (2023), Raffel and Chen (2023), Tang et al. (2023), Wang et al. (2023), Yan et al. (2023), Zhang et al. (2023a), Zhang and Feng (2023), Yang et al. (2024), Chen et al. (2024), Deng and Woodland (2024), Guo et al. (2024), Ko et al. (2024), Ma et al. (2024), Papi et al. (2024c), Papi et al. (2024a), Papi et al. (2024b), Tan et al. (2024), Zhang et al. (2024), Zhang and Feng (2024)

A.2.2 Cascade (49 papers)

Fügen et al. (2006a), Ryu et al. (2006), Fügen et al. (2007), Wolfel et al. (2008), Kolss et al. (2008), Fügen (2009), Cho et al. (2013), Fujita et al. (2013), Rangarajan Sridhar et al. (2013), Shimizu et al. (2013), Yarmohammadi et al. (2013), Oda et al. (2014), Wołk and Marasek (2014), Cho et al. (2015), Shavarani et al. (2015), Müller et al. (2016), Niehues et al. (2016), Wang et al. (2016), Cho et al. (2017), Dessloch et al. (2018), Siahbani et al. (2018), Wang et al. (2019), Xiong et al. (2019), Arivazhagan et al. (2020b), Arivazhagan et al. (2020a), Bahar et al. (2020), Elbayad et al. (2020a), Elbayad et al. (2020b), Iranzo-Sánchez et al. (2020), Macháček et al. (2020), Schneider and Waibel (2020), Wilken et al. (2020), Yao and Haddow (2020), Bahar et al. (2021), Bojar et al. (2021), Iranzo-Sánchez et al. (2021), Novitasari et al. (2021), Weller et al. (2021), Guo et al. (2022), Indurthi et al. (2022), Iranzo-Sánchez et al. (2022), Li et al. (2022), Sen et al. (2022), Subramanya and Niehues (2022), Wang et al. (2022b), Zeng et al. (2022), Guo et al. (2023), Iranzo-Sánchez et al. (2024), Guo et al. (2024)

A.3 By Presentation Strategy

A.3.1 Incremental (93 papers)

Ryu et al. (2006), Fügen et al. (2007), Wolfel et al. (2008), Kolss et al. (2008), Fügen (2009), Cho et al. (2013), Fujita et al. (2013), Rangarajan Sridhar et al. (2013), Shimizu et al. (2013), Yarmohammadi et al. (2013), Oda et al. (2014), Shavarani et al. (2015), Wang et al. (2016), Siahbani et al. (2018), Wang et al. (2019), Xiong et al. (2019), Arivazhagan et al. (2020a), Bahar et al. (2020), Elbayad et al. (2020a), Elbayad et al. (2020b), Han et al. (2020), Iranzo-Sánchez et al. (2020), Ma et al. (2020b), Ren et al. (2020), Schneider and Waibel (2020), Wilken et al. (2020), Nguyen et al. (2021a), Ma et al. (2021), Bahar et al. (2021), Chen et al. (2021), Iranzo-Sánchez et al. (2021), Karakanta et al. (2021), Liu et al. (2021b), Liu et al. (2021a), Nguyen et al. (2021b), Novitasari et al. (2021), Zaidi et al. (2021), Zeng et al. (2021), Chang and yi Lee (2022), Deng et al. (2022), Dong et al. (2022), Fukuda et al. (2022a), Gaido et al. (2022), Guo et al. (2022), Indurthi et al. (2022), Iranzo-Sánchez et al. (2022), Li et al. (2022), Papi et al. (2022a), Polák et al. (2022), Subramanya and Niehues (2022), Wang et al. (2022b), Xue et al. (2022), Zaidi et al. (2022), Zeng et al. (2022), Zhang et al. (2022), Zhang and Feng (2022), Zhu et al. (2022), Xue et al. (2023), Raffel et al. (2023), Barrault et al. (2023), Fu et al. (2023), Fukuda et al. (2023), Gaido et al. (2023), Guo et al. (2023), Huang et al. (2023), Iranzo-Sánchez et al. (2024), Ko et al. (2023), Ma et al. (2023), Papi et al. (2023d), Papi et al. (2023c), Papi et al. (2023b), Papi et al. (2023a), Polák (2023), Polák et al. (2023), Polák et al. (2023), Raffel and Chen (2023), Tang et al. (2023), Wang et al. (2023), Yan et al. (2023), Zhang et al. (2023a), Zhang and Feng (2023), Yang et al. (2024), Chen et al. (2024), Deng and Woodland (2024), Guo et al. (2024), Ko et al. (2024), Ma et al. (2024), Papi et al. (2024c), Papi et al. (2024a), Papi et al. (2024b), Tan et al. (2024), Zhang et al. (2024), Zhang and Feng (2024)

A.3.2 Re-translation (13)

Müller et al. (2016), Niehues et al. (2016), Arivazhagan et al. (2020b), Arivazhagan et al. (2020a), Macháček et al. (2020), Yao and Haddow (2020), Bojar et al. (2021), Weller et al. (2021), Amrhein and Haddow (2022), Sen et al. (2022), Omachi et al. (2023), Chen et al. (2023), Alastruey et al. (2023)

A.3.3 Undefined (5)

Fügen et al. (2006a), Wołk and Marasek (2014), Cho et al. (2015), Cho et al. (2017), Dessloch et al. (2018)

A.4 By Papers Mentioning Automatic Segmentation

A.4.1 Not Mentioned

Ryu et al. (2006), Fujita et al. (2013), Wołk and Marasek (2014), Cho et al. (2015), Cho et al. (2017), Dessloch et al. (2018), Siahbani et al. (2018), Xiong et al. (2019), Arivazhagan et al. (2020a), Bahar et al. (2020), Elbayad et al. (2020a), Elbayad et al. (2020b), Han et al. (2020), Ma et al. (2020b), Ren et al. (2020), Wilken et al. (2020), Yao and Haddow (2020), Nguyen et al. (2021a), Chen et al. (2021), Karakanta et al. (2021), Liu et al. (2021b), Nguyen et al. (2021b), Novitasari et al. (2021), Weller et al. (2021), Zaidi et al. (2021), Zeng et al. (2021), Chang and yi Lee (2022), Deng et al. (2022), Dong et al. (2022), Fukuda et al. (2022a), Guo et al. (2022), Indurthi et al. (2022), Iranzo-Sánchez et al. (2022), Papi et al. (2022a), Polák et al. (2022), Subramanya and Niehues (2022), Wang et al. (2022b), Xue et al. (2022), Zaidi et al. (2022), Zeng et al. (2022), Zhang et al. (2022), Zhang and Feng (2022), Zhu et al. (2022), Omachi et al. (2023), Chen et al. (2023), Xue et al. (2023), Raffel et al. (2023), Alastruey et al. (2023), Barrault et al. (2023), Fu et al. (2023), Fukuda et al. (2023), Gaido et al. (2023), Guo et al. (2023), Huang et al. (2023), Ko et al. (2023), Ma et al. (2023), Papi et al. (2023d), Papi et al. (2023c), Papi et al. (2023b), Papi et al. (2023a), Polák et al. (2023), Polák et al. (2023), Raffel and Chen (2023), Tang et al. (2023), Wang et al. (2023), Yan et al. (2023), Zhang et al. (2023a), Zhang and Feng (2023), Yang et al. (2024), Chen et al. (2024), Deng and Woodland (2024), Guo et al. (2024), Ko et al. (2024), Ma et al. (2024), Papi et al. (2024c), Papi et al. (2024a), Tan et al. (2024), Zhang et al. (2024), Zhang and Feng (2024)

A.4.2 Mentioned

Fügen et al. (2006a), Fügen et al. (2007), Wolfel et al. (2008), Kolss et al. (2008), Fügen (2009), Cho et al. (2013), Rangarajan Sridhar et al. (2013), Shimizu et al. (2013), Yarmohammadi et al. (2013), Oda et al. (2014), Shavarani et al. (2015), Müller et al. (2016), Niehues et al. (2016), Wang et al. (2016), Wang et al. (2019), Arivazhagan et al. (2020b), Iranzo-Sánchez et al. (2020), Macháček et al. (2020), Schneider and Waibel (2020), Ma et al. (2021), Bahar et al. (2021), Bojar et al. (2021), Iranzo-Sánchez et al. (2021), Liu et al. (2021a), Amrhein and Haddow (2022), Gaido et al. (2022), Li et al. (2022), Sen et al. (2022), Iranzo-Sánchez et al. (2024), Polák (2023), Papi et al. (2024b)