

SIMULSENSE: SENSE-DRIVEN INTERPRETING FOR EFFICIENT SIMULTANEOUS SPEECH TRANSLATION

Haotian Tan Hiroki Ouchi Sakriani Sakti

Nara Institute of Science and Technology, Japan

ABSTRACT

How to make human-interpreter-like read/write decisions for simultaneous speech translation (SimulST) systems? Current state-of-the-art systems formulate SimulST as a multi-turn dialogue task, requiring specialized interleaved training data and relying on computationally expensive large language model (LLM) inference for decision-making. In this paper, we propose SimulSense, a novel framework for SimulST that mimics human interpreters by continuously reading input speech and triggering write decisions to produce translation when a new sense unit is perceived. Experiments against two state-of-the-art baseline systems demonstrate that our proposed method achieves a superior quality-latency tradeoff and substantially improved real-time efficiency, where its decision-making is up to **9.6**× faster than the baselines.

Index Terms— simultaneous speech translation, LLM-based speech translation, decision policy, continuous integrate-and-fire

1. INTRODUCTION

Simultaneous speech translation (SimulST) is a challenging task to perform translation in real-time with low latency while maintaining high translation quality. A key component of such systems is the decision policy, which determines whether the currently available source context is sufficient for accurate translation or how many predicted tokens can be reliably displayed to users.

Early works on the SimulST have primarily relied on two paradigms: (1) designing dedicated training strategies to learn data-driven decision policies [1, 2], or (2) adapting pre-trained offline ST models with hand-crafted, rule-based decision policies [3, 4, 5]. However, these methods either require training multiple models for different latency scenarios or are highly dependent on the design of the rule-based decision policy. Recently, SimulST has been framed as a multi-turn dialogue task using Large Language Models (LLMs), where speech segments are treated as user prompts and translations as assistant responses [6, 7, 8]. To control the quality-latency trade-off, [8] employs a simple fixed decision policy that ensures translation begins only after a fixed number of source tokens have arrived. [7, 9] enable the LLM to learn a data-driven read/write policy that determines whether the accumulated source context is sufficient for translation. These multi-turn dialogue systems have achieved state-of-the-art performance in recent years. However, they require interleaved SimulST data for training, which is challenging to collect due to the high cost of manual annotation. [7, 9] propose prompting a powerful LLM to generate synthetic SimulST data, aiming to mimic human interpreter behavior and learn an optimal read/write policy. Nevertheless, their performance is highly dependent on the LLM capabilities and prompt engineering. [8] constructs training data from offline ST corpus using MFA and SimAlign while

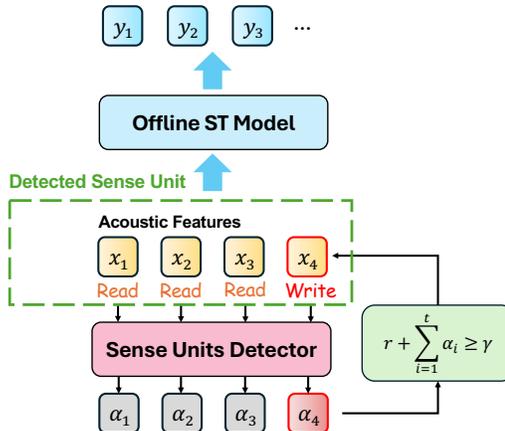


Fig. 1: The Sense Units Detector (SUD) assigns a weight to each acoustic feature of the input stream. When the accumulated weight sum (including a prior residual weight r) exceeds a triggering threshold γ , a sense unit is detected, which triggers the offline ST model to produce translations.

its decision policy does not mimic the human interpreters, making it sub-optimal.

In this paper, we propose SimulSense, a sense-aware framework for SimulST to achieve a better quality-latency trade-off. SimulSense is motivated by the cognitive process of human interpreters, where they attentively listen to the speaker’s speech and begin translation as soon as they perceive a sense unit (semantic chunk), which is the smallest linguistic unit capable of conveying a complete thought independently [10, 11]. This behavior enables low-latency, cognitively efficient, and semantically coherent output. To mimic this behavior, as illustrated in Figure 1, our approach continuously processes the input speech until a sense unit is detected by a dedicated Sense Units Detector (SUD) model. Upon detection, the write decision will be triggered to send the acoustic features to an offline ST model to produce translation. Despite [7, 9] also aiming to detect a sense unit to initiate translation, they require a full forward pass through the LLM to make decisions, which is inefficient even with the aid of the KV cache mechanism. In contrast, our method employs a lightweight SUD model that makes decisions directly from the audio encoder output, enabling efficient, real-time decision-making without costly LLM inference. Furthermore, prior works rely heavily on specialized interleaved training data, which may compromise translation quality. Our method instead leverages a well-established offline ST model trained on ground-truth parallel data, ensuring robust and high-quality translation capability. Our contributions are summarized as follows:

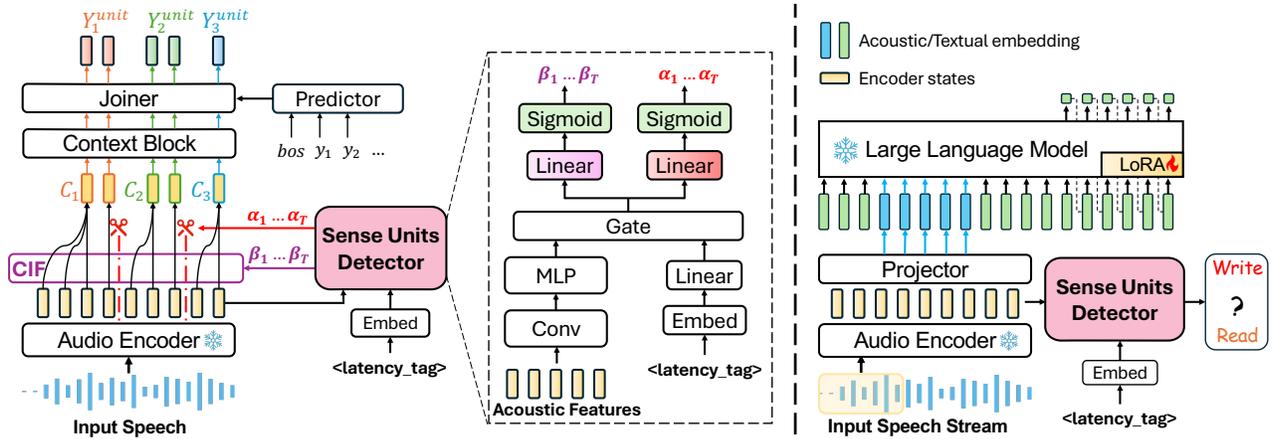


Fig. 2: The overview of our SimulSense framework. **Left:** A Sense-Aware Transducer (SAT) training pipeline that guides a lightweight Sense Units Detector (SUD) model in learning human-interpreter-like sense unit segmentation. **Right:** The SUD model dynamically perceives sense units from acoustic features and drives the LLM-based offline ST model to perform simultaneous translation.

- We propose SimulSense, a novel framework that mimics human interpreters to efficiently drive the simultaneous speech translation with dynamically detected sense units.
- We design a Sense Units Detection mechanism as well as its training pipeline, avoiding costly LLM inference and enabling efficient read/write decisions for SimulST.

2. RELATED WORK

2.1. Continuous Integrate-and-Fire for SimulST

Continuous Integrate-and-Fire (CIF) [12] is a soft and monotonic alignment mechanism originally proposed for ASR to learn the acoustic boundaries. The CIF operates in two modes: integration and firing. The integration mode uses a group of weights for speech frames to dynamically accumulate information until the sum of weights exceeds a certain threshold, at which point it enters the firing mode to produce an output for the accumulated information. In recent years, CIF has gained attention in the speech translation community. [13] integrated CIF into SimulST to locate acoustic boundaries, while [14, 15] employed it to decide how much input to read per translation token. Additionally, [16] used CIF to locate word boundaries to bring SimulST closer to text-to-text translation. Our work differs from these approaches. Instead of identifying acoustic or word boundaries, we aim to accumulate acoustic information corresponding to a semantically independent sense unit. Furthermore, we only use CIF to guide the learning of the Sense Units Detector in an ASR training pipeline. In our SimulST system, we only accumulate information without engaging in the integration and firing process.

2.2. Adapting Offline Systems for SimulST

Offline ST models consume the full source speech to achieve superior translation quality. Early works have explored leveraging offline models for SimulST through fixed, rule-based decision policies. [3] waits for k speech segments before the translation starts. [4] detects reliable model predictions to display to users by making an agreement between two consecutive chunks. [5] exploit the cross-attention mechanism to make such decisions. Unlike these fixed-

policy approaches, we aim to develop a flexible and adaptive decision policy that mimics human interpreter behavior.

3. PROPOSED METHOD

Figure 2 shows the overview of our SimulSense framework, which comprises a Sense Units Detector (SUD) for identifying semantic boundaries, a sense-aware transducer (SAT) training pipeline for optimizing read/write decisions, and an offline-trained speech translation model for high-quality translation generation.

3.1. Sense Units Data Creation

Inspired by recent work [9] that leverages LLMs for SimulST data creation, we create sense units data from existing speech-to-text datasets using a Qwen3-32B model. Given a speech-to-text sample (S, Y) , where S is the speech sequence and Y is the source transcription, we prompt the LLM to segment Y into N sense units $Y_{units} = (Y_1^{unit}, \dots, Y_N^{unit})$. Following [7], we use three latency settings (low, medium, high) that result in different segmentation densities: $N_{low} > N_{medium} > N_{high}$. Each sample is accompanied by a latency tag ($\langle low \rangle$, $\langle medium \rangle$, or $\langle high \rangle$) to enable latency-aware training.

3.2. Sense-Aware Transducer (SAT)

As illustrated in the left panel of Figure 2, the Sense-Aware Transducer (SAT) guides the SUD by integrating the CIF mechanism into a transducer-based ASR training pipeline. The SUD model functions as a weight predictor in the CIF mechanism. Unlike conventional CIF weight predictors, the SUD takes encoder output acoustic features $H = (h_1, \dots, h_T)$ and latency tag embeddings as inputs to predict two weight groups: $A = (\alpha_1, \dots, \alpha_T)$ and $B = (\beta_1, \dots, \beta_T)$ for each acoustic feature h_i , where T is the total number of frames. We perform CIF-like continuous weight accumulation using weights group A and dynamically segment the acoustic feature sequence into K non-overlapping segments $\{u_k\}_{k=1}^K$ by applying a threshold to the cumulative α -weights. Specifically, a segmentation boundary is triggered at frame t if the cumulative weights since the

last boundary exceed a triggering threshold γ , which is fixed to 1.0 during training:

$$r + \sum_{i=j+1}^t \alpha_i \geq \gamma, \quad (1)$$

where j denotes the frame index of the previous boundary, and $r \in [0, \gamma)$ is the residual weight from the prior segment. Upon triggering a boundary at frame t , the residual weight is updated as:

$$r \leftarrow r + \sum_{i=j+1}^t \alpha_i - \gamma, \quad (2)$$

This ensures that excess weight beyond the threshold is preserved for the next segment.

During training, given the sense units data (S, Y_{units}) containing N sense units, we apply the standard CIF scaling strategy to the weights group A to ensure:

$$\sum_{t=1}^T \alpha_t = N - 1. \quad (3)$$

This constraint guarantees exactly $N - 1$ boundary triggers, producing N acoustic feature segments $\{\mathcal{U}_k\}_{k=1}^N$ that correspond to the N sense units. We then generate unit-level integrated acoustic features by applying the CIF mechanism independently to each sense unit feature \mathcal{U}_k using weights group B and a fixed firing threshold λ :

$$C_k = \text{CIF}(\mathcal{U}_k; B, \lambda = 1.0), \quad (4)$$

where $C_k = (c_k^1, \dots, c_k^{M_k})$ represents the CIF features for the k -th sense unit and M_k denotes the feature length. Given the target transcription $Y_k^{unit} = (y_k^1, \dots, y_k^{L_k})$ for the k -th unit, where L_k is the number of target tokens, we apply the scaling strategy to weights B before CIF to enforce $M_k = L_k$. This ensures one-to-one alignment between the CIF features $C_k = (c_k^1, \dots, c_k^{M_k})$ and target tokens, such that $c_k^j \leftrightarrow y_k^j$. Following [17], we append a context block after the CIF module to enhance contextual dependencies among the CIF outputs.

This process yields length-aligned joiner inputs $(\{C_k\}_{k=1}^N, Y)$ and produces three-dimensional output logits. Unlike traditional RNN-T training, we employ cross-entropy loss \mathcal{L}_{Joint} for more efficient optimization. To force the model to learn alignment between acoustic feature segments and sense units, we introduce two quantity losses. First, we constrain the total number of acoustic feature segments:

$$\mathcal{L}_{Qua1} = \left| \sum_{t=1}^T \alpha_t - N \right|. \quad (5)$$

Second, we ensure that CIF outputs match the target token count for each sense unit:

$$\mathcal{L}_{Qua2} = \sum_{k=1}^N \left| \sum_{i=1}^{M_k} \beta_k^i - L_k \right|. \quad (6)$$

Additionally, we apply cross-entropy loss \mathcal{L}_{LM} to the predictor model to enhance contextual semantic understanding. The total training loss combines all components:

$$\mathcal{L}_{Total} = \mathcal{L}_{Joint} + \mathcal{L}_{Qua1} + \mathcal{L}_{Qua2} + \mathcal{L}_{LM}. \quad (7)$$

3.3. Read/Write Decision Policy

We employ an offline ST model to perform translation, which shares the same pre-trained acoustic encoder as the SAT network. During inference, as illustrated in the right panel of Figure 2, we construct the SimulST pipeline by integrating the pre-trained SUD model with the offline ST model. The audio encoder processes the input speech stream in a chunk-by-chunk manner. Given a latency tag, the SUD model dynamically weights the acoustic features and triggers the offline ST model to generate translation when a new sense unit is detected (i.e., when Equation 1 is satisfied).

4. EXPERIMENTAL SETUP

4.1. Data

The International Conference on Spoken Language Translation (IWSLT) is an annual conference and evaluation campaign that promotes research on spoken language translation. In this paper, we follow the data condition of the IWSLT 2025 Simultaneous Track [18] to use CoVoST-2 [19] as the training set and ACL 60/60 [20] for both validation and evaluation. We conduct experiments on three translation directions: English-to-German (En→De), English-to-Japanese (En→Ja), and English-to-Chinese (En→Zh).

4.2. Model Architecture

We use Whisper-large-v3 [21] as the audio encoder. In the SAT network, a Conformer encoder serves as the context block, followed by a UGBP joiner [22]. The predictor is Qwen-3-0.6B, fine-tuned with LoRA during SAT training. The SUD model comprises a two-layer convolution block, a two-layer MLP, and a gating mechanism. Our offline ST model stacks the Whisper encoder, a two-layer linear projector, and the Qwen-3-8B LLM. During training, we freeze the Whisper encoder and finetune the LLM with LoRA.

4.3. Baselines

We compare our proposed method against two state-of-the-art SimulST systems.

NAIST-2025 [23]: This model adapts an LLM-based offline ST system for SimulST by applying a Local Agreement (LA) policy [4]. At IWSLT 2025 [18], it was ranked 1st for the En→Ja language pair and 2nd for both En→Zh and En→De in the low-latency setting.

Dialogue-LLM: This baseline is built upon the methodology described in [7], utilizing the multi-turn dialogue mechanism of an LLM. To ensure a fair comparison, we use the same audio encoder, projector, LLM, and CovoST-2-derived training data as our SimulSense framework. We show later that the Dialogue-LLM baseline outperforms the NAIST-2025 system in terms of quality-latency tradeoff in Section 5.1.

4.4. Evaluation Setup

We use BLEU (\uparrow) to evaluate the translation quality. Latency is measured using the LAAL metric [24]. We report the computation-aware latency with all evaluations performed on an NVIDIA RTX A6000 GPU. To further assess the efficiency of the read/write decision policy across systems, we report the average inference time per decision and the Real-Time Factor (RTF). We use different triggering thresholds $\gamma = \{0.5, 1.0, \dots, 5.0\}$ to control the quality-latency tradeoff of our proposed method.

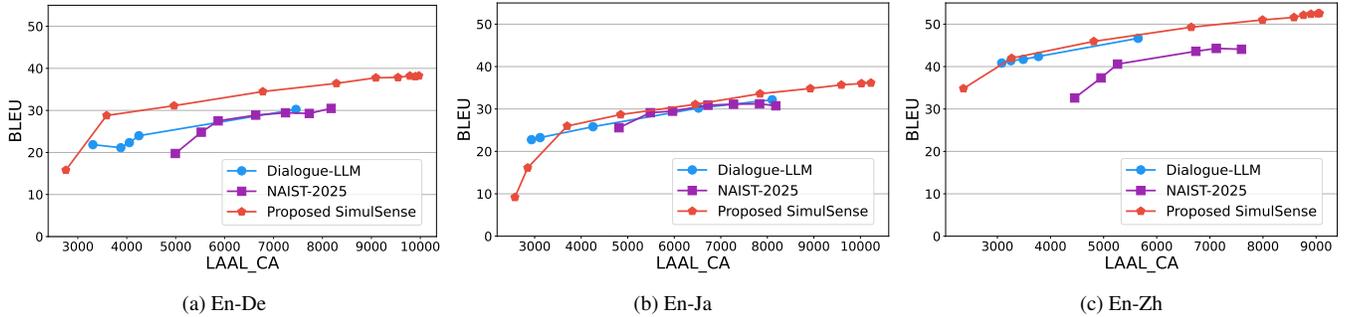


Fig. 3: Quality-latency trade-off of our SimulSense framework compared to two state-of-the-art SimulST systems on the En→De, En→Ja, and En→Zh directions of the ACL 60/60 test set. SimulSense achieves a superior quality-latency tradeoff across all language pairs.

5. RESULTS

5.1. Quality-Latency Tradeoff

Figure 3 illustrates the computation-aware quality-latency tradeoff of our SimulSense framework against the baselines. Our system significantly outperforms both baselines for the En→De language pair, achieving a maximum BLEU improvement of 6.9 against Dialogue-LLM (at around 3.5s latency) and 11.4 against NAIST-2025 (at around 5.0s latency). Furthermore, it also consistently outperforms both baselines for the En→Zh direction. For the En→Ja pair, the translation quality of our system surpasses that of the baselines when latency exceeds 3 seconds. Notably, the proposed method is capable of achieving a wider latency span, ranging from approximately 2.5s to 10s, by simply adjusting the SUD triggering threshold. This flexibility makes it highly suitable for real-world deployment.

5.2. Efficiency on Read/Write Decision Policy

Unlike NAIST-2025 and Dialogue-LLM, which derive their decisions from LLM outputs, our SimulSense framework employ a more efficient decision module comprising an audio encoder and the SUD model. As shown in Table 1, this decision module achieves a significantly lower average inference time. Specifically, it is $3.0\times$ faster than Dialogue-LLM and $9.6\times$ faster than NAIST-2025 per decision. Moreover, its RTF is $8.1\times$ and $11.3\times$ lower than those of the two baselines, respectively, indicating a substantial improvement in real-time efficiency.

Table 1: Efficiency comparison of the decision policies across different systems.

Model	Avg. Inference Time (ms)	RTF
NAIST-2025	371.3	0.180
Dialogue-LLM	116.2	0.130
Proposed SimulSense	38.6	0.016

5.3. Impact of Latency Tags

We also investigated the use of latency tags to control the quality-latency tradeoff. However, as shown in Figure 4, our results suggest that these tags primarily constrain the maximum latency and do not significantly impact the overall tradeoff. Consequently, we used a high-latency tag for all experiments.

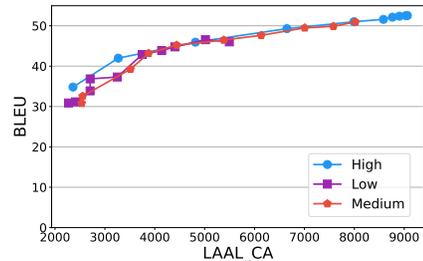


Fig. 4: The impact of using different latency tags.

5.4. WER of Sense-Aware Transducer

As shown in Table 2, we evaluated three SAT variants with varying numbers of context block layers and attention heads. We selected SAT-Small as the final architecture, as it achieved a superior WER of 67.7% on the ACL 60/60 test set. Interestingly, although the SAT training yielded a relatively high WER, the SUD model was still able to make effective decisions for the SimulST task. This may be because during SAT training, the SUD model effectively learned to identify sense unit boundaries but did not perform well at transcribing the unit-level acoustic information integrated by the CIF mechanism.

Table 2: WER comparison across SAT variants with varying context block layers and attention heads.

Model	Context Block	Layer	Heads	#Params (M)	WER (%)
SAT-Mini	×	—	—	41	74.5
SAT-Small	✓	2	2	100	67.7
SAT-Large	✓	8	4	232	69.1

6. CONCLUSION

We proposed SimulSense, a sense-aware framework that enables SimulST to make human-interpreter-like read/write decisions. Unlike existing methods that rely on LLM inference, SimulSense employs a highly efficient Sense Units Detector (SUD) model for decision-making and leverages an offline-trained speech translation model to conduct high-quality translation. Experiments against two state-of-the-art SimulST systems across three translation directions demonstrate that SimulSense achieves a superior quality-latency tradeoff and substantially improved real-time efficiency.

7. ACKNOWLEDGMENTS

Part of this work is supported by JSPS KAKENHI Grant Numbers JP21H05054 and JP23K21681.

8. REFERENCES

- [1] Xutai Ma, Juan Pino, et al., “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*, Dec. 2020, pp. 582–587.
- [2] Xingshan Zeng, Liangyou Li, et al., “RealTranS: End-to-end simultaneous speech translation with convolutional weighted-shrinking transformer,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, Aug. 2021, pp. 2461–2474.
- [3] Mingbo Ma, Liang Huang, et al., “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*, July 2019, pp. 3025–3036.
- [4] Danni Liu, Gerasimos Spanakis, and Jan Niehues, “Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis Selection,” in *Proc. Interspeech*, 2020, pp. 3620–3624.
- [5] Sara Papi, Marco Turchi, et al., “AlignAtt: Using Attention-based Audio-Translation Alignments as a Guide for Simultaneous Speech Translation,” in *Proc. INTERSPEECH*, 2023, pp. 3974–3978.
- [6] Minghan Wang, Thuy-Trang Vu, et al., “Conversational SimulMT: Efficient simultaneous translation with large language models,” in *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*. July 2025, pp. 93–105, Association for Computational Linguistics.
- [7] Biao Fu, Donglei Yu, et al., “Efficient and adaptive simultaneous speech translation with fully unidirectional architecture,” *arXiv preprint arXiv:2504.11809*, 2025.
- [8] Siqi Ouyang, Xi Xu, and Lei Li, “InfiniSST: Simultaneous translation of unbounded speech with large language model,” in *Findings of the Association for Computational Linguistics: ACL 2025*. July 2025, pp. 3032–3046, Association for Computational Linguistics.
- [9] Shanbo Cheng, Zhichao Huang, et al., “Towards achieving human parity on end-to-end simultaneous speech translation via llm agent,” *arXiv preprint arXiv:2407.21646*, 2024.
- [10] Marianne Lederer, “Simultaneous interpretation—units of meaning and other features,” in *Language interpretation and communication*, pp. 323–332. Springer, 1978.
- [11] Roderick Jones, *Conference interpreting explained*, Routledge, 2014.
- [12] Linhao Dong and Bo Xu, “Cif: Continuous integrate-and-fire for end-to-end speech recognition,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6079–6083.
- [13] Qian Dong, Yaoming Zhu, et al., “Learning when to translate for streaming speech,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, May 2022, pp. 680–694.
- [14] Chih-Chiang Chang and Hung-yi Lee, “Exploring Continuous Integrate-and-Fire for Adaptive Simultaneous Speech Translation,” in *Interspeech 2022*, 2022, pp. 5175–5179.
- [15] Keqi Deng and Phil Woodland, “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)*, 2024, pp. 8235–8251.
- [16] Keqi Deng, Wenxi Chen, et al., “Simul2S-LLM: Unlocking simultaneous inference of speech LLMs for speech-to-speech translation,” in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2025, pp. 16718–16734.
- [17] Tian-Hao Zhang, Dinghao Zhou, et al., “Cif-t: A novel cif-based transducer architecture for automatic speech recognition,” in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10531–10535.
- [18] Idris Abdulmumin, Victor Agostinelli, et al., “Findings of the IWSLT 2025 evaluation campaign,” in *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*. July 2025, pp. 412–481, Association for Computational Linguistics.
- [19] Changhan Wang, Juan Pino, et al., “CoVoST: A diverse multilingual speech-to-text translation corpus,” in *Proceedings of The 12th Language Resources and Evaluation Conference*. May 2020, pp. 4197–4203, European Language Resources Association.
- [20] Elizabeth Salesky, Kareem Darwish, et al., “Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology,” in *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. July 2023, pp. 62–78, Association for Computational Linguistics.
- [21] Alec Radford, Jong Wook Kim, et al., “Robust speech recognition via large-scale weak supervision,” in *Proceedings of the 40th International Conference on Machine Learning*. 23–29 Jul 2023, vol. 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518, PMLR.
- [22] Chao Zhang, Bo Li, et al., “Improving the fusion of acoustic and text representations in rnn-t,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8117–8121.
- [23] Haotian Tan, Ruhayah Faradishi Widiaputri, et al., “NAIST simultaneous speech translation system for IWSLT 2025,” in *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*. July 2025, pp. 369–378, Association for Computational Linguistics.
- [24] Sara Papi, Marco Gaido, et al., “Over-generation cannot be rewarded: Length-adaptive average lagging for simultaneous speech translation,” in *Proceedings of the Third Workshop on Automatic Simultaneous Translation*, July 2022, pp. 12–17.