Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Empowering Large Language Models for End-to-End Speech Translation Leveraging Synthetic Data

Yu Pu[1,*], *Xiaoqian Liu*[2,*], *Guangyu Zhang*[1], *Zheng Yan*[2], *Wei-Qiang Zhang*[1,#], *Xie Chen*[3,#]

[1]Department of Electronic Engineering, Tsinghua University, China
[2]School of Comput. Sci. and Eng., Northeastern University, China
[3]Dept. of Comput. Sci. and Eng., Shanghai Jiao Tong University, China

puy23@mails.tsinghua.edu.cn, wqzhang@tsinghua.edu.cn

Abstract

Speech-to-speech translation (S2ST) is a key technology for seamless cross-lingual communication. Traditional cascaded systems, which involve speech recognition, text translation, and speech synthesis, are prone to error propagation and latency. In this work, we present SLAM-TR, an end-to-end speech translation model which directly maps input speech to output speech, eliminating the need for intermediate text representations. By fine-tuning from the large language model Qwen2-0.5B, SLAM-TR achieves superior performance over the cascaded baseline and state-of-the-art open-source models with minimal training time. Additionally, SLAM-TR demonstrates strong generalization, achieving an ASR-BLEU score of 8.2 on the FLEURS benchmark, outperforming both cascaded and open-source systems. Addressing the challenge of limited natural speech translation data, we propose SynStard-1000, a 1,000-hour synthetic speech translation dataset. We create the dataset by synthesizing speech data from high-quality parallel text pairs. And the open-source release of this dataset contributes to the field of speech-to-speech translation.

Index Terms: speech translation, end-to-end system, large language model, data synthesis

1. Introduction

Speech-to-speech translation (S2ST) is a critical task that plays an essential role in bridging communication gaps between speakers of different languages and regions [1]. Traditional approaches to speech-to-speech translation generally rely on a cascaded system composed of automatic speech recognition (ASR), machine translation (MT), and text-to-speech synthesis (TTS) [2, 3, 4, 5]. With the advent of large language models (LLMs), the quality of text-based translation has significantly improved [6, 7, 8, 9], indirectly enhancing the performance of cascaded S2ST systems. However, these cascaded systems suffer from inherent limitations, including slower processing speeds due to the sequential nature of the pipeline [10, 11, 12, 13, 14] and potential information loss during intermediate text representations [15, 16, 17, 18, 19].

End-to-end speech translation systems offer a promising solution to these challenges, as they can process audio input and generate output speech in a single, unified model [20, 21, 22, 23, 24]. Radford et al. (2023) [6] proposed Whisper, which leverages a transformer-based architecture to achieve speech-to-text translation. Seamless Communication et al. (2023) [7] introduced SeamlessM4T, which integrates speech recognition, translation, and synthesis into a unified end-to-end system, enabling robust performance in multilingual speech translation tasks. Despite their potential, training such models from scratch without leveraging the power of large language models remains a major hurdle. End-to-end systems, by design, require a significant amount of parallel speech translation data, which is relatively scarce compared to textual data.

In parallel, recent advancements in end-to-end speech dialogue systems have successfully adapted large language models to directly process and generate speech, demonstrating the feasibility of fine-tuning LLMs for end-to-end speech tasks. Xie et al. (2024) [25] discretized speech using SNAC encoding and decoded the text tokens and speech tokens in parallel within the large language model to implement Mini-Omni, an end-to-end speech dialogue system. Fang et al. (2024) [26] put forward LLaMA-Omni, an end-to-end speech dialogue system that can generate a speech response in 226 ms. In this work, we investigate the adaptation of this approach to the domain of speech-to-speech translation. Specifically, we explore whether the framework of end-to-end multimodal large language models can be leveraged to achieve high-quality Chinese-English speech translation.

To overcome the data scarcity issue, we first constructed a 1000-hour speech translation dataset by utilizing a speech synthesis system to generate spoken versions of available text translation data. This synthetic data was then used to fine-tune an end-to-end speech translation model from the pre-trained LLM. The main content of this paper is summarized in Figure 1. Our experiments validate the applicability of this approach, underscoring its potential to advance the field of end-to-end speech translation.

2. SynStard-1000 Construction

We propose the SynStard-1000 (**Syn**thetic **S**peech-to-Speech **Tra**nslation **D**ataset) for end-to-end speech translation. This dataset is derived from parallel text pairs in WMT News Commentary v18 [27], and we used the F5-TTS [28] model to synthesize the corresponding speech. The voice prompts for synthesis were sourced from the Seed-TTS-Eval [29] dataset.

2.1. Parallel Corpus Collection and Cleaning

We select the WMT News Commentary v18 [27] dataset as the primary source of parallel text data. The dataset contains over 400,000 sentence pairs in English and Chinese, covering a wide range of news topics and providing high linguistic diversity and complexity. To ensure data quality, we perform the following cleaning operations on the raw data. First, we remove sentence pairs containing non-standard characters (e.g.,

[*]These authors contributed equally.
[#]Corresponding author.

This work was supported by the National Natural Science Foundation of China under Grant No. 62276153.

10.21437/Interspeech.2025-2341

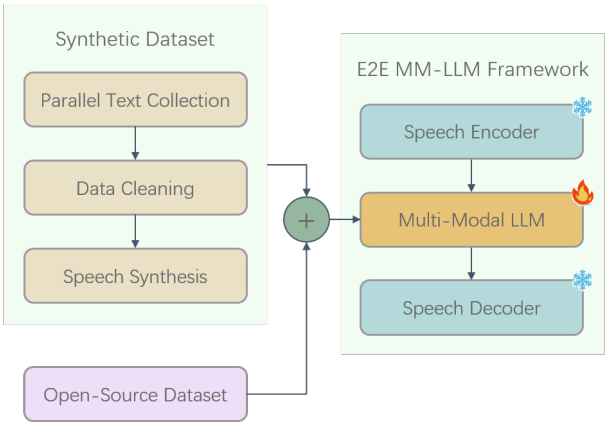


Figure 1: *Flowgram of the synthetic dataset construction and end-to-end speech translation model training.*

garbled text, emojis). Second, we filter out excessively long or short sentence pairs by removing the top and bottom 5% of pairs based on the ratio of Chinese to English word counts. This step aims to eliminate samples of extreme length that could interfere with model training.

After cleaning, we obtained approximately 390,000 high-quality parallel sentence pairs, providing a solid foundation for the English-Chinese speech translation task.

2.2. Speech Synthesis

To convert text data into speech data, we choose the open-source F5-TTS [28] model as our speech synthesis tool. F5-TTS is a flow-matching-based text-to-speech (TTS) model developed by the Shanghai Jiao Tong University team. With its fast inference speed (RTF of 0.15) and high-quality speech generation capabilities, F5-TTS is an ideal choice for data synthesis. We used 200 Chinese prompts from the Seed-TTS-Eval [29] dataset as reference audios for F5-TTS. The Seed-TTS-Eval dataset is an open-source speech synthesis evaluation dataset released by ByteDance, containing diverse speech samples suitable for multi-scenario speech generation tasks.

Ultimately, we constructed the SynStard-1000 dataset comprising approximately 390,000 sentence pairs, totaling around 1,000 hours of speech. This dataset has been published on HuggingFace[1]. Alongside the natural speech translation datasets, it will provide reliable data support for training and evaluating end-to-end speech translation models.

3. S2ST Model Training

In this section, we describe the training process of our end-to-end speech translation model SLAM-TR (**Tr**anslation model based on **SLAM**-Omni Framework). In the following, we detail the architecture, components, and training methodology of SLAM-TR.

3.1. Training Framework

SLAM-TR is trained based on the SLAM-Omni [30] framework. SLAM-Omni is a single-stage training framework that supports controllable voice characteristics in speech dialogue systems. It uses an audio encoder to extract speech features, then concurrently generates speech and auxiliary text tokens within a multimodal large language model. Finally, a vocoder is used to decode the audio tokens into speech. A simplified architecture diagram of SLAM-Omni is shown in Figure 2. SLAM-Omni achieves superior performance with minimal training resources, requiring only 15 hours of training on 4 GPUs.

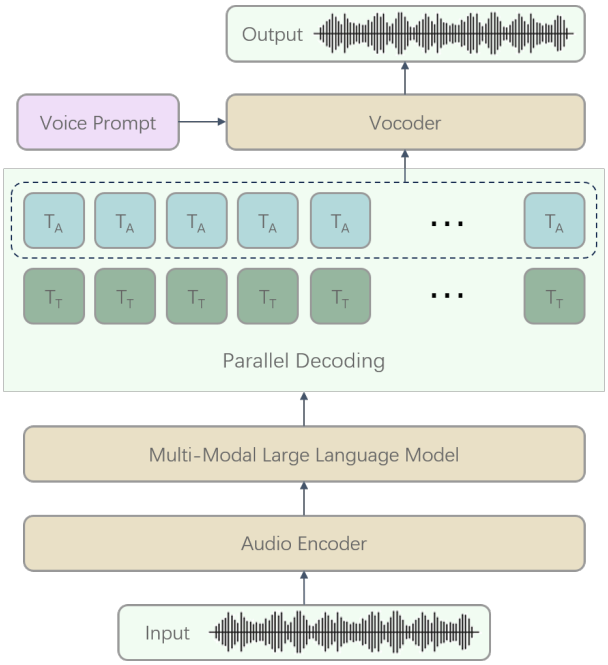


Figure 2: *Simplified architecture diagram of SLAM-Omni, where TA represents the audio token and TT represents the text token.*

3.1.1. Speech Encoder

We employ Whisper [6], an open-source automatic speech recognition (ASR) model developed by OpenAI, as our speech encoder. Whisper is trained on 680,000 hours of multilingual and multitask supervised data, enabling robust performance across various languages and accents. We discard the decoder of Whisper and use it solely for extracting speech features, skipping the transcription of text, thereby forming the frontend of our end-to-end model.

3.1.2. Base Model

We adopt Qwen2-0.5B as our base model. Qwen2 [31] is a lightweight yet powerful large language model optimized for efficient inference and training. It supports a wide range of natural language processing tasks, including translation, summarization, and question-answering. With 0.5 billion parameters, Qwen2-0.5B strikes a balance between performance and computational cost.

3.1.3. Vocoder

For speech decoding, we use CosyVoice [32], a high-quality text-to-speech (TTS) model known for its natural and fluent output. CosyVoice supports multilingual speech synthesis and allows fine-grained control over voice characteristics, such as pitch, speed, and emotion. We use the decoder of CosyVoice to convert audio tokens into speech, rather than starting from text.

[1] https://huggingface.co/datasets/cksqs/SynStard-1000

3.2. Data

In this study, we utilize two datasets to train SLAM-TR: the SynStard-1000 dataset, which we have constructed, and the Seamless-Align dataset. Seamless-Align [7], proposed by Meta, is an open-source speech translation dataset widely used in related research. It contains multilingual audio data aligned across over 30 languages. For our experiments, we specifically used the Chinese-English subset of Seamless-Align. Because the Seamless-Align dataset contains only audio data, we employ Whisper-small [6] to generate transcriptions as reference texts for translation accuracy assessment. The information of the training datasets is shown in Table 1.

Table 1: *Datasets for Model Training*

Dataset	Number of Pairs	Duration (h)	
		Chinese	English
SynStard-1000	393k	1.08k	1.28k
Seamless-Align	482k	1.18k	1.25k

To evaluate the performance of SLAM-TR, we employ both internal validation subsets derived from our training data and an external benchmark dataset. We created validation sets by extracting 0.1% of the data from SynStard-1000 and Seamless-Align, respectively, with the remaining data reserved for training. In addition, we utilize the FLEURS (Few-shot Learning Evaluation of Universal Representations of Speech) [33] dataset, a multilingual dataset designed for evaluating speech recognition, translation, and synthesis systems. We specifically use the Chinese-to-English test subset of FLEURS, containing 349 pairs of speech. The information of the evaluation datasets is shown in Table 2.

Table 2: *Datasets for Model Evaluation*

Dataset	Number of Pairs
SynStard-1000	394
Seamless-Align	482
FLEURS	349

4. Experimental Settings

4.1. Baseline

To provide a comprehensive comparison with our end-to-end speech translation model SLAM-TR, we construct a cascaded baseline system. This system follows the traditional pipeline approach, which consists of three distinct stages: speech recognition, text translation, and speech synthesis. We employ Whisper-small to transcribe the Chinese speech into text. The transcribed text is then passed to Qwen2-0.5B, which serves as the text translation module. Finally, the translated text is synthesized into English speech using CosyVoice. This cascaded architecture mirrors the components corresponding to the speech encoder, base LLM, and vocoder of SLAM-TR, respectively.

To ensure a fair comparison, we fine-tune the Qwen2-0.5B model on the same training dataset used for training SLAM-TR. The fine-tuning process involved full-parameter optimization, with hyperparameters identical to those of SLAM-TR. By fine-tuning Qwen2-0.5B on the same dataset, we ensure that the cascaded system benefits from the same domain-specific knowledge as SLAM-TR.

4.2. Evaluation Metrics

To assess the performance of our end-to-end speech translation model, we adopt the ASR-X evaluation framework, which means transcribing the output speech into text using an ASR model and evaluating the transcribed text against reference translations using text-based translation metrics. We use Whisper-small as the ASR component and four different text-based translation metrics as the "X" component in ASR-X. Detailed text-based translation metrics are introduced in Table 3.

Table 3: *Details of Text-Based Evaluation Metrics*

Text Metrics	Target
BLEU	precision of n-gram overlaps
ROUGE-1	overlap of unigrams (single words)
METEOR	extended BLEU considering synonym matching, stemming, and word order
Sent-Transformer	semantic similarity based on pre-trained transformer models

4.3. Training Details

We train the SLAM-TR model on four NVIDIA A800 GPUs. It takes only 7.5 hours to perform full-parameter fine-tuning from the base 0.5B model on the SynStard-1000 and Seamless-Align datasets for 3 epochs. The hyperparameters during training are set as Table 4.

Table 4: *Hyperparameters during SLAM-TR Training*

learning rate	1e-4
batch size	8
training precision	fp16
optimization method	Adam
warmup steps	3,000

5. Results

5.1. Evaluation on Internal Datasets

5.1.1. Natural Dataset Seamless-Align

In this section, we present the experimental results of our end-to-end speech translation model, SLAM-TR, on the Seamless-Align dataset. We compare its performance against several baseline systems, including the cascaded baseline and two variants of the Whisper-CosyVoice pipeline, using Whisper-small (denoted as Ws) and Whisper-medium (denoted as Wm). The Whisper-CosyVoice pipeline translates speech to text using Whisper, followed by speech synthesis with CosyVoice. Although neither of these components has been fine-tuned on our

training data, both are state-of-the-art models in the field and remain highly competitive for speech translation tasks.

Table 5: *Result of Evaluation on Seamless-Align*

Model	BLEU	ROUGE-1	METEOR	Sent-T
SLAM-TR	7.24	0.25	0.18	0.46
Cascaded	6.74	0.22	0.16	0.39
Ws + Cosy	3.59	0.16	0.12	0.31
Wm + Cosy	5.25	0.20	0.15	0.39

As shown in the Table 5, SLAM-TR outperforms the cascaded baseline system across all evaluation metrics. The result indicates that the end-to-end approach is more effective in preserving semantic fidelity. This aligns with our hypothesis that eliminating the bottlenecks of cascaded systems can lead to improved performance.

It is worth noting that the ASR-X evaluation framework introduces additional noise due to ASR transcription errors, which results in lower absolute metric values compared to traditional text-based translation tasks. However, the relative differences between models remain consistent and meaningful, providing a reliable basis for comparing their performance.

5.1.2. Synthetic Dataset SynStard-1000

The results of the experiment on SynStard-1000 are shown in Table 6. Despite fine-tuning the text translation component of the cascaded system on the SynStard-1000 dataset, the cascaded model performs poorly, achieving an ASR-BLEU score of only 2.12. This is significantly lower than the performance of SLAM-TR. The poor performance of the cascaded system can be attributed to the lower naturalness of the synthetic speech in the dataset, which introduces substantial errors during the automatic ASR stage. These errors propagate through the pipeline, leading to degraded translation quality.

Table 6: *Result of Evaluation on SynStard-1000*

Model	BLEU	ROUGE-1	METEOR	Sent-T
SLAM-TR	9.96	0.30	0.24	0.54
Cascaded	2.12	0.11	0.07	0.16
Ws + Cosy	2.00	0.10	0.06	0.14
Wm + Cosy	1.99	0.10	0.06	0.13

The open-source Whisper-CosyVoice systems also exhibit poor performance on the SynStard-1000 dataset, which underscores the challenges faced by cascaded systems in scenarios with suboptimal speech quality.

In contrast to the cascaded systems, SLAM-TR achieves significantly higher scores across all metrics. This demonstrates that eliminating intermediate text representations and directly modeling the relationship between input and output speech features can effectively mitigate error propagation. The end-to-end approach is more robust to variations in speech quality, making it suitable for a wider range of scenarios.

5.2. Evaluation on External Benchmark

To ensure a fair and comprehensive evaluation, we conduct additional experiments on the FLEURS benchmark dataset, as internal datasets may not provide an equitable assessment of open-source models. The FLEURS dataset provides a more generalized setting for assessing model performance instead.

Table 7: *Result of Evaluation on FLEURS*

Model	BLEU	ROUGE-1	METEOR	Sent-T
SLAM-TR	8.20	0.30	0.24	0.55
Cascaded	2.95	0.14	0.09	0.17
Ws + Cosy	6.99	0.24	0.19	0.37
Wm + Cosy	7.51	0.26	0.21	0.41

The evaluation result is shown in Table 7. SLAM-TR demonstrates significantly better performance than the cascaded system, indicating that fine-tuning an end-to-end multi-modal LLM yields superior generalization compared to fine-tuning a text-based LLM alone. Additionally, SLAM-TR outperforms the open-source model Whisper-CosyVoice on external datasets, providing objective evidence of its strong generalization capability across diverse speech translation tasks.

In summary, SLAM-TR consistently outperforms both state-of-the-art open-source systems and fine-tuned cascaded baselines across internal and external evaluation datasets. All systems adopt the same speech encoder and speech decoder, and the cascaded baselines are fine-tuned using the same training data and hyperparameters as SLAM-TR. Therefore, the performance gain of SLAM-TR can be attributed solely to the advantages of end-to-end speech translation modeling. These results empirically validate that the end-to-end architecture effectively reduces error accumulation and improves performance in speech-to-speech translation tasks.

6. Discussion

In this work, we present SynStard-1000, a synthetic speech translation dataset. We create the SynStard-1000 dataset by synthesizing speech data from high-quality parallel text pairs using the F5-TTS model. This approach enables us to generate a large-scale, Chinese-English speech translation dataset, addressing the challenge of limited natural speech translation data. The open-source release of this dataset contributes to the advancement of speech-to-speech translation.

Based on the natural and synthetic speech translation data, we present SLAM-TR, an end-to-end speech translation model. We demonstrate the superior performance of SLAM-TR across multiple datasets. Unlike cascaded systems, which are highly sensitive to variations in speech quality, SLAM-TR adapts seamlessly to diverse evaluation scenarios. The robustness of SLAM-TR highlights the advantages of eliminating intermediate text representations and directly modeling the relationship between input and output speech features. This approach not only mitigates error propagation but also enhances generalization to diverse and challenging speech translation tasks.

Our work contributes to the growing body of research on end-to-end speech translation, offering a scalable and efficient solution for breaking down language barriers in real-world scenarios. Due to computational and time constraints, our experiments have focused on Chinese-to-English speech translation. Future work will aim to expand our experiments to additional language pairs, especially those with less available natural speech translation data. By exploring a broader range of languages, we plan to further assess the generalization capabilities of SLAM-TR and enhance its applicability to a wider variety of real-world speech translation tasks.

7. References

[1] M. Sarim, S. Shakeel, L. Javed, M. Nadeem *et al.*, "Direct speech to speech translation: A review," *arXiv preprint arXiv:2503.04799*, 2025.

[2] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The ATR multilingual speech-to-speech translation system," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 2, pp. 365–376, 2006.

[3] L. Arya, A. Roy Chowdhury, and S. R. M. Prasanna, "Direct vs cascaded speech-to-speech translation using transformer," in *Speech Comput.*, 2023, pp. 258–270.

[4] S. Zhang, Q. Fang, S. Guo, Z. Ma, M. Zhang, and Y. Feng, "StreamSpeech: Simultaneous speech-to-speech translation with multi-task learning," in *Findings Assoc. Comput. Linguistics 2024*, vol. 1, 2024, pp. 8964–8986.

[5] H. Zhang, N. Si, Y. Chen, X. Yang, D. Qu, and W.-Q. Zhang, "Improving speech translation by cross-modal multi-grained contrastive learning," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 1075–1086, 2023.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. Mach. Learn.*, vol. 202, 2023, pp. 28 492–28 518.

[7] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman *et al.*, "SeamlessM4T: Massively multilingual & multimodal machine translation," *arXiv preprint arXiv:2308.11596*, 2023.

[8] D. Gao, H. Wang, Y. Li, X. Sun, Y. Qian, B. Ding, and J. Zhou, "Text-to-SQL empowered by large language models: A benchmark evaluation," in *Proc. VLDB Endow.*, vol. 17, no. 5, 2024, pp. 1132–1145.

[9] X. Chen, S. Zhang, Q. Bai, K. Chen, and S. Nakamura, "LLaST: Improved end-to-end speech translation system leveraged by large language models," in *Findings Assoc. Comput. Linguistics 2024*, vol. 1, 2024, pp. 6976–6987.

[10] L. Bentivogli, M. Cettolo, M. Gaido, A. Karakanta, A. Martinelli, M. Negri, and M. Turchi, "Cascade versus direct speech translation: Do the differences still make a difference?" in *Findings Assoc. Comput. Linguistics 2021*, vol. 1, 2021, pp. 2873–2887.

[11] T. Etchegoyhen, H. Arzelus, H. Gete, A. Alvarez, I. G. Torre, J. M. Martín-Doñas, A. González-Docasal, and E. B. Fernandez, "Cascade or direct speech translation? A case study," *Appl. Sci.*, vol. 12, no. 3, 2022, art. no. 1097.

[12] T. K. Lam, S. Schamoni, and S. Riezler, "Cascaded models with cyclic feedback for direct speech translation," in *Proc. ICASSP*, 2021, pp. 7508–7512.

[13] P. Bahar, T. Bieschke, R. Schlüter, and H. Ney, "Tight integrated end-to-end training for cascaded speech translation," in *Proc. 2021 IEEE Spoken Lang. Tech. Workshop*, 2021, pp. 950–957.

[14] P. S. K. Paleti, B. Saluru, S. Vatturi, B. Yohoshiva, and N. P. Challa, "Cascaded speech-to-speech translation system focusing on low-resource Indic languages," in *Proc. 2024 3rd Ed. IEEE Delhi Sec. Flagship Conf.*, 2024.

[15] G. Zhou, T. K. Lam, A. Birch, and B. Haddow, "Prosody in cascade and direct speech-to-text translation: A case study on Korean wh-phrases," in *Findings Assoc. Comput. Linguistics 2024*, vol. 1, 2024, pp. 674–683.

[16] R. Dabre and H. Song, "NICT's cascaded and end-to-end speech translation systems using Whisper and IndicTrans2 for the Indic task," in *Proc. 21st Int. Conf. Spoken Lang. Transl.*, 2024, pp. 17–22.

[17] V. A. K. Tran, D. Thulke, Y. Gao, C. Herold, and H. Ney, "Does joint training really help cascaded speech translation?" in *Proc. 2022 Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 4480–4487.

[18] C. Le, Y. Qian, D. Wang, L. Zhou, S. Liu, X. Wang, M. Yousefi, Y. Qian, J. Li, and M. Zeng, "Transvip: Speech to speech translation system with voice and isochrony preservation," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 89 682–89 705, 2024.

[19] Y. Ren, J. Liu, X. Tan, C. Zhang, T. Qin, Z. Zhao, and T.-Y. Liu, "SimulSpeech: End-to-end simultaneous speech to text translation," in *Findings Assoc. Comput. Linguistics 2025*, vol. 1, 2020, pp. 3787–3796.

[20] A. Lee, P.-J. Chen, C. Wang, J. Gu, S. Popuri, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang, J. Pino, and W.-N. Hsu, "Direct speech-to-speech translation with discrete units," in *Findings Assoc. Comput. Linguistics 2022*, vol. 1, 2022, pp. 3327–3339.

[21] C. Huber, T. A. Dinh, C. Mullov, N.-Q. Pham, T. B. Nguyen, F. Retkowski, S. Constantin, E. Ugan, D. Liu, Z. Li, S. Koneru, J. Niehues, and A. Waibel, "End-to-end evaluation for low-latency simultaneous speech translation," in *Proc. 2023 Conf. Empirical Methods Natural Lang. Process. Syst. Demonstrations*, 2023, pp. 12–20.

[22] H. Inaguma, K. Duh, T. Kawahara, and S. Watanabe, "Multilingual end-to-end speech translation," in *2019 IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 570–577.

[23] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," in *Proc. ICASSP*, 2018, pp. 6224–6228.

[24] P. Bahar, P. Wilken, T. Alkhouli, A. Guta, P. Golik, E. Matusov, and C. Herold, "Start-before-end and end-to-end: Neural speech translation by AppTek and RWTH Aachen University," in *Proc. 17th Int. Conf. Spoken Lang. Transl.*, 2020, pp. 44–54.

[25] Z. Xie and C. Wu, "Mini-Omni: Language models can hear, talk while thinking in streaming," *arXiv preprint arXiv:2408.16725*, 2024.

[26] Q. Fang, S. Guo, Y. Zhou, Z. Ma, S. Zhang, and Y. Feng, "LLaMA-Omni: Seamless speech interaction with large language models," *arXiv preprint arXiv:2409.06666*, 2024.

[27] N. I. Nikolov, Y. Hu, M. X. Tan, and R. H. Hahnloser, "Character-level Chinese-English translation through ASCII encoding," in *Proc. 3rd Conf. Mach. Transl. Res. Papers*, 2018, pp. 10–16.

[28] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, "F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching," *arXiv preprint arXiv:2410.06885*, 2024.

[29] P. Anastassiou, J. Chen, J. Chen, Y. Chen, Z. Chen, Z. Chen, J. Cong, L. Deng, C. Ding, L. Gao *et al.*, "Seed-TTS: A family of high-quality versatile speech generation models," *arXiv preprint arXiv:2406.02430*, 2024.

[30] W. Chen, Z. Ma, R. Yan, Y. Liang, X. Li, R. Xu, Z. Niu, Y. Zhu, Y. Yang, Z. Liu *et al.*, "SLAM-Omni: Timbre-controllable voice interaction system with single-stage training," *arXiv preprint arXiv:2412.15649*, 2024.

[31] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[32] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma *et al.*, "CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *arXiv preprint arXiv:2407.05407*, 2024.

[33] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "FLEURS: Few-shot learning evaluation of universal representations of speech," in *Proc. 2022 IEEE Spoken Lang. Tech. Workshop*, 2023, pp. 798–805.