

Non-autoregressive speech translation with understanding, translation, reordering and LLM-augmented correction

Yu-Chen Kuan and Kuan-Yu Chen

*Department of Computer Science and Information Engineering,
National Taiwan University of Science and Technology, Taipei, Taiwan*

APSIPA
Transactions on
Signal and
Information
Processing

309

Received 29 September 2025
Revised 26 January 2026
9 March 2026
Accepted 12 March 2026

Abstract

Non-autoregressive speech translation (NAR-ST) offers low latency through parallel decoding but often falls short of autoregressive (AR) systems in producing fluent and well-structured translations. The authors present UTRo-NAST, a novel NAR-ST framework that decomposes speech translation into three consecutive subtasks: source speech understanding, word-by-word translation and target-side reordering. This divide-and-conquer design enhances interpretability and training stability while preserving fully parallel inference. To further improve performance, the authors introduce a plug-and-play large language models (LLM)-augmented post-correction strategy that refines UTRo-NAST outputs with prompting mechanisms. Experiments on the MuST-C benchmark across eight language pairs show that UTRo-NAST consistently outperforms existing NAR-ST models and delivers translation quality comparable to strong AR baselines, while maintaining faster decoding. The LLM-augmented variant achieves competitive or superior results compared to recent LLM-integrated ST systems, offering a practical path to stronger NAR-ST without iterative refinement or costly LLM fine-tuning. Overall, these results demonstrate the effectiveness and scalability of UTRo-NAST for practical speech translation.

Keywords Speech translation, Non-autoregressive modeling, Large language models, Postcorrection, CTC

Paper type Research paper

1. Introduction

With the rapid globalization of communication, the demand for speech translation (ST) has grown substantially. Traditional ST systems are typically built on cascaded architectures that integrate automatic speech recognition (ASR) and machine translation (MT). While effective, these systems suffer from error propagation and high latency. End-to-end speech translation (E2E-ST) has emerged as a more efficient alternative, removing intermediate steps to reduce latency and improve overall performance (Inaguma *et al.*, 2020; Wang *et al.*, 2020b; Weiss *et al.*, 2017a; Zhao *et al.*, 2021). Current E2E-ST models generally follow either an autoregressive (AR) or a non-autoregressive (NAR) decoding paradigm. AR models achieve strong translation quality by generating tokens sequentially, but at the cost of high inference latency (Fang *et al.*, 2022; Gao *et al.*, 2024; Ye *et al.*, 2022, 2021; Zhou *et al.*, 2023). By contrast, NAR models decode in parallel, substantially reducing latency and



© Yu-Chen Kuan and Kuan-Yu Chen. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at <http://creativecommons.org/licences/by/4.0/>

APSIPA Transactions on Signal
and Information Processing
Vol. 15 No. 1, 2026
pp. 309-331
Emerald Publishing Limited
2048-7703
DOI 10.1108/ATSP-09-2025-0097

making them attractive for real-time applications (Inaguma *et al.*, 2021a; Inaguma *et al.*, 2021b; Xu *et al.*, 2023). However, their inability to effectively handle word-order and syntactic differences across languages often leads to degraded fluency, particularly for structurally divergent language pairs.

Recent advances in NAR E2E-ST have largely centered on two decoding paradigms: connectionist temporal classification (CTC) (Graves *et al.*, 2006) and conditional masked language modeling (CMLM) (Ghazvininejad *et al.*, 2019). CTC-based methods directly generate output sequences from speech without requiring presegmentation or explicit length prediction (Xu *et al.*, 2023). By introducing a blank symbol, CTC enables constant-time, fully parallel decoding. Nevertheless, the conditional independence assumption underlying CTC prevents the model from capturing contextual dependencies during inference, thereby limiting translation quality. CMLM-based methods address this issue by adopting an iterative refinement strategy, typically implemented with the mask-predict algorithm (Inaguma *et al.*, 2021a, 2021b). These models repeatedly mask and regenerate low-confidence tokens to improve output quality. While CMLM alleviates the independence assumption of CTC, it relies on an auxiliary length predictor and multiple decoding iterations, which undermines efficiency compared to single-pass NAR methods.

Building a fully integrated NAR E2E-ST model is inherently complex, difficult to debug, and challenging to optimize due to the tight coupling between components (Omachi *et al.*, 2023; Ran *et al.*, 2021). Motivated by a divide-and-conquer philosophy, we propose decomposing the ST process into simpler, more interpretable subtasks. To this end, we introduce UTRo-NAST, a novel NAR-ST framework comprising an acoustic encoder, a word-by-word mapping encoder and a reordering decoder, corresponding to the subtasks of source speech understanding, word-level translation and target-side reordering (Kuan and Chen, 2025). Each component is optimized with task-specific objectives under a multitask learning scheme, enabling the system to progressively transform source speech into coherent target-language output.

With the rapid progress of large language models (LLMs), many state-of-the-art ST systems have integrated them to enhance translation fluency and robustness. Building on these advances, we extend UTRo-NAST with a plug-and-play LLM-augmented correction strategy. To leverage the strengths of LLMs while maintaining a lightweight and efficient design, we introduce three prompting schemes that further improve translation performance.

We evaluate UTRo-NAST on the MuST-C benchmark (Di Gangi *et al.*, 2019), which spans eight language pairs. Experimental results show that UTRo-NAST consistently outperforms strong NAR-ST baselines (Inaguma *et al.*, 2021a; Xu *et al.*, 2023). UTRo-NAST achieves an average BLEU improvement of +1.02 across all language pairs, compared to the best-performing NAR model. Moreover, relative to AR systems, UTRo-NAST provides a $3.97\times$ speedup in inference while maintaining competitive translation quality. Finally, incorporating LLM-augmented correction, the UTRo-NAST variants deliver competitive or superior performance relative to recent LLM-enhanced ST systems, thereby validating the effectiveness and scalability of the framework for real-world speech translation applications.

2. Related work

2.1 Speech-to-text translation

Speech-to-text translation (ST) aims to directly translate spoken utterances in a source language into text in a target language. Early approaches relied on pipeline architectures, in which automatic speech recognition and machine translation systems were trained independently and cascaded sequentially (Stentford and Steer, 1990). While modular and effective, pipeline models suffer from error propagation, as ASR errors inevitably degrade

translation quality. To overcome this limitation, researchers explored end-to-end ST (E2E-ST) models that map speech directly to translated text without intermediate transcription. Pioneering works demonstrated the feasibility of end-to-end training using sequence-to-sequence architectures with attention mechanisms (Bérard *et al.*, 2016; Weiss *et al.*, 2017b). These models eliminate intermediate text representations and enable joint optimization, though they typically require large-scale data. Depending on the decoding paradigm, E2E-ST systems can be broadly categorized into autoregressive (AR) and non-autoregressive (NAR) architectures.

2.1.1 Autoregressive modeling. Autoregressive ST models achieve strong performance by generating each output token conditioned on previously generated tokens (Inaguma *et al.*, 2020; Wang *et al.*, 2020b; Zhao *et al.*, 2021). This sequential dependency allows the model to capture rich context and perform accurate language modeling. The typical formulation of AR-based ST models relies on the conditional probability decomposition:

$$P(Z|X) = \prod_{u=1}^U P(z_u | z_{<u}, X), \quad (1)$$

where X denotes the input speech features and $Z = (z_1, z_2, \dots, z_U)$ represents the target translation sequence.

Recent studies have further advanced AR-based ST models through multitask learning strategies (Chen *et al.*, 2024c; Cheng *et al.*, 2023; Fang and Feng, 2023; Fang *et al.*, 2022; Gao *et al.*, 2024; Tian *et al.*, 2025; Tsiamas *et al.*, 2024b; Wang *et al.*, 2024; Yan *et al.*, 2024; Ye *et al.*, 2022, 2021; Zhang *et al.*, 2024; Zhou *et al.*, 2023), which jointly optimize ST, ASR, and MT by sharing encoders and/or decoders, thereby enhancing model generalization. Xu *et al.* (2024) and Yan *et al.* (2023) further combined CTC with cross-entropy loss to improve alignment and accuracy. While these techniques enhance translation quality and robustness, the sequential nature of autoregressive decoding still limits inference speed.

2.1.2 Non-autoregressive modeling. In contrast, NAR models reduce latency by generating tokens in parallel, making them attractive for real-time use cases (Gu *et al.*, 2018). The basic NAR formulation assumes conditional independence among target tokens:

$$P(Z|X) = \prod_{u=1}^U P(z_u | X). \quad (2)$$

Here, each token z_u depends only on the input speech X , without reference to other target tokens. This assumption makes it difficult to model token regularity and syntactic differences, often degrading fluency for structurally divergent languages.

Recent work on NAR E2E-ST can be categorized into two paradigms:

- (1) purely CTC-based approaches; and
- (2) conditional masked language modeling (CMLM) with iterative refinement.

Inaguma *et al.* introduced the Orthros series (Inaguma *et al.*, 2021a, 2021b), a flexible framework supporting both CTC and CMLM decoders. Chuang *et al.* (2021) systematically examined reordering in CTC-based models and proposed a multitask strategy that jointly optimizes source and target CTC losses to improve alignment. More recently, Xu *et al.* presented CTC-NAST (Xu *et al.*, 2023), which incorporates intermediate CTC supervision, prediction-aware encoding, and curriculum mixup to further improve robustness and translation quality.

2.2 Speech translation with large language models

Recent advances in ST increasingly integrate LLMs into end-to-end architectures. Given their proven capacity for language understanding and generation, LLMs are being leveraged to address persistent limitations in ST, particularly in contextual reasoning and fluency. Zhang *et al.* proposed LST (Zhang *et al.*, 2023a), a joint fine-tuning framework that combines speech-pretrained models with LLMs via prompt-based adaptation, yielding notable quality gains. Building on this idea, LLaST (Chen *et al.*, 2024a) embeds LLMs directly into the E2E-ST pipeline and uses a Dual-LoRA training scheme. In parallel, SALM (Chen *et al.*, 2024e) introduced a speech-augmented prompting paradigm, in which acoustic representations are fused with task-specific instructions as LLM prompts. This design enables in-context learning and zero-shot generalization across diverse speech-language modalities.

Beyond integrated pipelines, other methods enhance ST with LLMs via postprocessing or auxiliary training. GenTranslate (Hu *et al.*, 2024) leverages N-best hypotheses to construct semantically consistent outputs with LLM support. Dou *et al.* (2025) refined translations through postdecoding prompts that jointly improve ASR and ST outputs. TDR (Zhou *et al.*, 2025) improves context modeling through multitask learning across ASR, MT and CoT-based ST, with Jensen–Shannon divergence to enforce consistency. CoT-Prompting (Hu *et al.*, 2025) further extends CoT reasoning by integrating speech encodings, ASR hypotheses and fixed prompts to strengthen semantic inference. Collectively, these approaches highlight the growing role of LLMs in advancing ST performance.

2.3 Summary

In summary, AR-based ST models deliver strong translation quality through sequential decoding but are constrained by high inference latency, limiting their applicability in real-time scenarios. NAR-based approaches address this limitation with parallel decoding, yet often struggle to maintain fluency and accurately model cross-lingual reordering. Recent efforts have attempted to mitigate these issues through CTC and CMLM variants, but challenges remain in balancing efficiency with translation quality. At the same time, LLMs have opened new directions for improving fluency, contextual reasoning and robustness, either through integration into end-to-end pipelines or as postprocessing modules. These developments highlight both the promise and limitations of existing approaches, motivating our proposed UTRo-NAST framework, which combines modular NAR modeling with an LLM-augmented correction process to achieve efficient and high-quality speech translation.

3. Proposed methodology

In this study, we propose a novel framework called UTRo-NAST, short for Understanding, Translation and Reordering for Non-Autoregressive Speech Translation. The core idea of UTRo-NAST is to decompose the end-to-end speech translation (E2E-ST) task into three explicitly defined, consecutive subtasks: source speech understanding, word-level translation and target-side reordering. To this end, UTRo-NAST is composed of three specialized components: an acoustic encoder, a word-by-word mapping encoder and a reordering decoder. By adopting a modular design with a multitask learning strategy, UTRo-NAST improves interpretability, enhances modeling flexibility, and boosts overall translation performance, while preserving the inference efficiency characteristic of NAR models. The complete architecture is illustrated in Figure 1.

Beyond the core NAR framework, we introduce a plug-and-play LLM-augmented correction strategy that refines the preliminary translations generated by UTRo-NAST. This component leverages three prompting variants: text-only prompting, ASR-enhanced

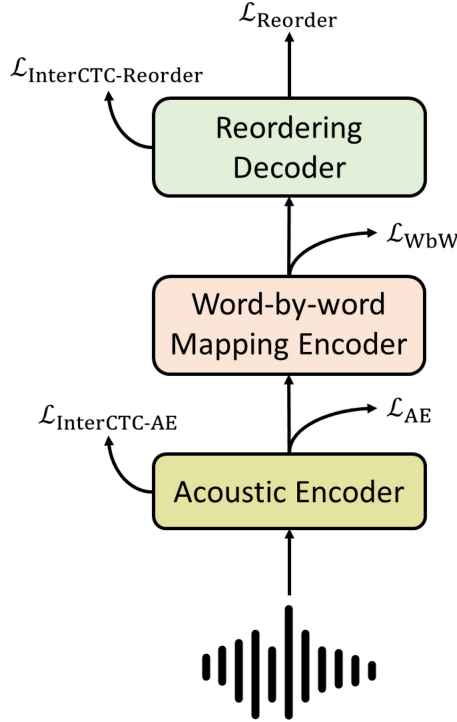


Figure 1. Overall architecture of the proposed UTRo-NAST framework

prompting and ASR + chain-of-thought prompting. These progressively incorporate additional contextual grounding and reasoning – ranging from direct refinement of translations, to conditioning on ASR hypotheses, to stepwise error identification and correction – thereby improving fluency and grammaticality without compromising efficiency. Together, the NAR backbone and the LLM-augmented correction strategy form a comprehensive framework that balances speed, interpretability and translation quality.

3.1 Acoustic encoder

Given a speech translation data set, each training example is represented as a triplet (X, Y, Z) , where $X = (x_1, \dots, x_T)$ denotes the input acoustic feature sequence (e.g. log-Mel filterbanks), $Y = (y_1, \dots, y_S)$ is the corresponding source-language transcription, and $Z = (z_1, \dots, z_U)$ is the target-language translation.

To capture both local temporal patterns and global dependencies in the acoustic sequence X , we adopt a stack of Conformer layers (Gulati *et al.*, 2020) to construct the encoder:

$$h^{(l)} = \text{Conformer}^{(l)}(h^{(l-1)}), \quad l = 1, \dots, L, \quad (3)$$

where $h^{(l)}$ denotes the hidden representation at the l -th layer and $h^{(0)} = X$. The final output $h^{(L)}$ serves as input to the downstream translation module.

To explicitly preserve the semantic content of the source speech, we use the source-language transcription Y as an auxiliary supervision signal, optimizing the encoder with a CTC loss:

$$\mathcal{L}_{\text{AE}} = -\log P_{\text{CTC}}(Y|h^{(L)}). \quad (4)$$

Furthermore, we incorporate two auxiliary techniques, Intermediate CTC (InterCTC) (Lee and Watanabe, 2021) and Prediction-Aware Encoding (PAE) (Nozaki and Komatsu, 2021; Xu et al., 2023), to mitigate the conditional independence assumption of CTC and improve training stability at selected intermediate layers. InterCTC facilitates gradient flow and encourages transcription-aware features by applying CTC loss to intermediate layers:

$$\mathcal{L}_{\text{InterCTC}}^{(l)} = -\log P_{\text{CTC}}(Y|h^{(l)}). \quad (5)$$

In parallel, PAE injects token-level prediction feedback into intermediate features. Specifically, a softmax distribution $P_{\text{CTC}}(h^{(l)})$ is projected back to the hidden space via a linear transformation and added to the original representation:

$$\tilde{h}^{(l)} = h^{(l)} + \text{Linear}(P_{\text{CTC}}(h^{(l)})), \quad (6)$$

where $\text{Linear}(\cdot)$ maps the token distribution to the hidden dimension. The feedback-enhanced representation $\tilde{h}^{(l)}$ is then passed to the next Conformer block.

3.2 Word-by-word mapping encoder

After extracting acoustic features, the next step is to translate the source speech content into the target language. To simplify this process, we design this stage to operate in a monotonic, word-by-word manner.

To bridge the modality gap between acoustic and textual representations, we apply a linear projection to map the acoustic embeddings $h^{(L)}$ into a latent space aligned with token embeddings:

$$\hat{h} = \text{Linear}(h^{(L)}). \quad (7)$$

The projected sequence $\hat{h}$ is then processed by a stack of sequence modeling blocks (e.g. Transformer or Conformer):

$$\bar{h}^{(k)} = f^{(k)}(\bar{h}^{(k-1)}), \quad k = 1, \dots, K, \quad (8)$$

where $f^{(k)}$ denotes the k -th sequence modeling layer, and $\bar{h}^{(0)} = \hat{h}$.

To realize word-by-word translation, the encoder is trained to produce a rough target sequence Z' that preserves the source sentence word order (see Section 4.1 and Table 1). Although Z' may not conform to correct target-language syntax, it serves as an intermediate objective that encourages the model to learn a monotonic word-level translation mapping. The training objective is defined using a CTC loss:

$$\mathcal{L}_{\text{wbw}} = -\log P_{\text{CTC}}(Z'|\bar{h}^{(K)}), \quad (9)$$

where $\bar{h}^{(K)}$ denotes the final representation from the sequence modeling stack.

Table 1. An example illustrating the word-by-word mapping and the final reordered translation in the UTRo-NAST framework

Source text	I am going to talk today about energy and climate
Word-by-word mapping	Ich bin gehen Zu sprechen Heute um Energie Und Klima
Final Translation	Heute spreche ich zu Ihnen über Energie und Klima

3.3 Reordering decoder

Building on the preliminary word-by-word translation, the final stage reorders the intermediate sequence into a fluent and grammatically correct sentence in the target language. To this end, the reordering decoder takes $\bar{h}^{(K)}$ as input and applies a stack of decoder layers:

$$\tilde{h}^{(j)} = g^{(j)}(\tilde{h}^{(j-1)}, \bar{h}^{(K)}), j = 1, \dots, J, \quad (10)$$

where $g^{(j)}$ denotes the j th decoder block and $\tilde{h}^{(0)} = \bar{h}^{(K)}$. Each decoder layer incorporates a cross-attention mechanism, using $\tilde{h}^{(j-1)}$ as the query and $\bar{h}^{(K)}$ as the key and value, thereby guiding the reordering process with global semantic context.

The decoder is optimized with a CTC loss defined over the ground-truth target sequence Z :

$$\mathcal{L}_{\text{Reorder}} = -\log P_{\text{CTC}}(Z|\tilde{h}^{(J)}). \quad (11)$$

As in the earlier modules, InterCTC and PAE are applied at selected intermediate layers to provide auxiliary semantic supervision and further enhance reordering quality.

3.4 The understanding, translation and reordering for non-autoregressive speech translation framework

In summary, UTRo-NAST decomposes the end-to-end speech translation task into three consecutive subtasks: source speech understanding, word-level translation and target-side reordering. Each subtask is handled by a dedicated module with its own training objective and the entire framework is optimized end-to-end with a multitask loss:

$$\begin{aligned} \mathcal{L} = & \mathcal{L}_{\text{AE}} + \mathcal{L}_{\text{WbW}} + \mathcal{L}_{\text{Reorder}} \\ & + \frac{1}{M} \sum_{m=1}^M \mathcal{L}_{\text{InterCTC-AE}}^{(m)} \\ & + \frac{1}{N} \sum_{n=1}^N \mathcal{L}_{\text{InterCTC-Reorder}}^{(n)} \end{aligned} \quad (12)$$

where $\mathcal{L}_{\text{InterCTC-AE}}^{(m)}$ denotes the auxiliary InterCTC loss at the m -th selected layer of the acoustic encoder, with M being the total number of such layers. Similarly, $\mathcal{L}_{\text{InterCTC-Reorder}}^{(n)}$ and N represent the corresponding InterCTC losses and number of selected layers in the reordering decoder.

During inference, UTRo-NAST performs fully parallel decoding. The final translation is generated by the reordering decoder via greedy decoding over the CTC outputs, followed by the removal of blank symbols and repeated tokens. This procedure enables fast and efficient non-autoregressive decoding while maintaining high translation quality.

3.5 Large language model-augmented correction strategy

To further enhance performance, we leverage the power of LLMs through a postcorrection process. This approach provides a lightweight and effective alternative to fully end-to-end LLM architectures: it requires no joint training, incurs minimal engineering overhead and can be flexibly enabled or disabled at inference time.

Our design is motivated by two key considerations. First, directly fine-tuning or tightly coupling LLMs with ST models is computationally expensive and typically demands large amounts of supervised data, which restricts practical deployment. Second, recent studies have shown that carefully crafted prompts can elicit strong correction capabilities from off-the-shelf LLMs in zero- or few-shot settings (Dou et al., 2025). Based on these insights, we decouple generation and correction: the ST backbone produces a preliminary translation, and an external LLM refines it via prompting. This decoupled design allows flexible substitution of different pretrained LLMs without re-engineering the ST backbone.

To harness LLMs efficiently, we design a three-tier prompting framework, as illustrated in Table 2:

Table 2. Proposed prompting strategies for LLM-augmented correction, illustrated with the English–German (en–de) language pair. The three variants include: *TP* (Text-Only prompting), *AP* (ASR-Enhanced prompting) and *ACoT* (ASR + chain-of-thought prompting)

Prompt Template (En-De example)
Text-only prompting (TP) You are a German expert translator that will be tasked with translating and improving the spelling, grammar and literary quality of a piece of text. Please rewrite the translated text to be more faithful and fluent. Please execute the command for the following sentence: German translated text: <#Translation Hyp#>
ASR-enhanced prompting (AP) You are a German expert translator that will be tasked with translating and improving the spelling, grammar and literary quality of a piece of text. Please rewrite the translated text to be more faithful and fluent according to the given English transcription. Please execute the command for the following sentence: English transcription: <#Transcription Hyp#> German translated text: <#Translation Hyp#>
ASR + Chain-of-Thought Prompting (ACoT) As an expert translation post editor, your task is to improve the German translation for the below English text. English: <#Transcription Hyp#> German: <#Translation Hyp#> To accomplish this, follow these steps: Step 1: Proposed Improvements – brainstorm and design improvements that make the German translation more faithful and fluent according to the given English transcription text. Step 2: Improved Translation – output the German translation with proposed improvements that increase translation faithfulness and fluency according to the given English transcription text. The output must be surrounded with < > in one column.

- (1) *Text-Only Prompting (TP)* – the LLM receives only the preliminary translation and is instructed to refine it as a translation expert.
- (2) *ASR-Enhanced Prompting (AP)* – in addition to the preliminary translation, the ASR recognition result of the source speech is provided, anchoring corrections to the spoken content and reducing semantic drift.
- (3) *ASR + Chain-of-Thought Prompting (ACoT)* – building upon AP, the LLM is further guided through a stepwise reasoning process: first identifying potential errors, then proposing a refined translation.

These three levels of prompting (TP → AP → ACoT) progressively increase contextual grounding and reasoning support, thereby enabling LLM-augmented correction to flexibly balance simplicity, robustness, and translation accuracy, while highlighting its scalability and adaptability within real-world speech translation pipelines.

4. Experimental setup

4.1 Data set and preprocessing

We conduct experiments on the MuST-C v1 corpus (Di Gangi *et al.*, 2019), a multilingual ST data set derived from TED talks. It covers eight translation directions from English (En) to German (De), Spanish (Es), French (Fr), Italian (It), Dutch (Nl), Portuguese (Pt), Romanian (Ro) and Russian (Ru). All models are trained in a bilingual setting, with model selection based on the development set and final results reported on the `test - COMMON` test set for each language pair.

To construct supervision signals for the word-by-word mapping encoder, we generate word-level translation targets using bilingual dictionaries obtained via Google Translate. Specifically, all unique English words in the training set are extracted and individually translated into the target language. Each English sentence is then converted into a word-level translation by replacing its words with their dictionary translations while preserving the original order. This procedure is applied uniformly across all language pairs to ensure consistency. An illustrative example is shown in Table 1.

We acknowledge that this dictionary-based construction may introduce noise, particularly for polysemous words and context-dependent lexical choices. However, these word-level targets are intended only as a rough, monotonic intermediate signal for learning cross-lingual correspondences while preserving source-side ordering. The subsequent reordering decoder incorporates sentence-level context to refine the intermediate sequence and generate fluent target sentences, which is expected to alleviate the impact of local imperfections in the word-by-word supervision. In addition, our framework is not tied to a specific dictionary source and can readily incorporate alternative lexicons or alignment-based mappings as stronger intermediate supervision, which we leave for future work.

Preprocessing follows the standard Fairseq pipeline (Wang *et al.*, 2020b). Utterances shorter than 5 frames or longer than 3,000 frames are discarded. Acoustic features are extracted as 80-dimensional log-Mel filterbanks with a 25 ms window and 10 ms stride. SentencePiece is applied for tokenization, using a shared vocabulary of 10,000 subwords across source and target languages to support cross-lingual modeling. To improve robustness, we apply SpecAugment (Park *et al.*, 2019), while speed perturbation is not used.

4.2 Training configuration

Our models are implemented in Fairseq (Wang *et al.*, 2020b) and optimized with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.98$) under the default learning rate schedule. Dropout is set to 0.15 during training.

To control complexity and isolate architectural effects, we adopt a consistent setup for the sequence modeling blocks $f(\cdot)$ and $g(\cdot)$ used in the word-by-word mapping encoder and reordering decoder. We evaluate two variants: *UTRo-NAST-T*, where both components use Transformer blocks, and *UTRo-NAST-C*, where both use Conformer blocks. This comparison highlights architectural differences without introducing the confounding effects of hybrid configurations.

The acoustic encoder, word-by-word mapping encoder and reordering decoder contain $L = 12$, $K = 3$, and $J = 9$ layers, respectively. Each layer consists of 512 hidden units, 8 attention heads and a feed-forward dimension of 2,048.

To stabilize training and improve representation learning, InterCTC and PAE are applied at layers six and nine of the acoustic encoder, and at layers three and six of the word-by-word encoder. For the reordering decoder, a curriculum mix-up strategy is integrated into PAE with a fixed ratio of 0.8 (Qian et al., 2021; Wang et al., 2020a; Xu et al., 2023).

All models are trained on four NVIDIA RTX 3090 GPUs with a maximum batch size of 20,000 tokens. Gradient accumulation is performed with an update frequency of four. Training runs for up to 100 epochs per language pair, with early stopping based on development BLEU and a patience of 20 epochs. During inference, the final model is obtained by averaging the top ten checkpoints ranked by development set BLEU. Evaluation is conducted with case-sensitive SacreBLEU (Post, 2018).

5. Experiments

5.1 Main results

We first evaluate the proposed UTRo-NAST framework under two architectural variants: UTRo-NAST-T and UTRo-NAST-C. Results are reported with and without sequence-level knowledge distillation (Seq-KD) (Kim and Rush, 2016), where the distilled translations are obtained from an autoregressive MT model with a beam size of five. The overall results are summarized in the bottom section of Table 3.

Without Seq-KD, UTRo-NAST-C consistently outperforms UTRo-NAST-T, achieving an average BLEU score of 29.2 versus 28.6. This advantage is particularly notable for syntactically flexible languages such as Romanian (25.1 vs 24.1) and Russian (18.6 vs 17.5), suggesting that Conformer-based modeling better captures long-range dependencies and structural variation. With Seq-KD applied, both variants improve further: UTRo-NAST-C + Seq-KD reaches 29.7 BLEU, while UTRo-NAST-T + Seq-KD achieves 29.6 BLEU. The reduced gap indicates that knowledge distillation mitigates architectural differences by providing smoother and more deterministic training targets.

5.2 Comparison with representative non-autoregressive-speech translation models

We next compare UTRo-NAST with representative non-autoregressive ST (NAR-ST) baselines, including Orthros-CTC (Inaguma et al., 2021a, 2021b), Orthros-CMLM (Inaguma et al., 2021a, 2021b), CTC-NAST (Xu et al., 2023), Hard Multitask CTC (Yan et al., 2024) and NAST-S2T (Ma et al., 2024). The results are summarized in the lower section of Table 3. Overall, UTRo-NAST achieves stronger translation performance across most language pairs. Among the compared NAR baselines, CTC-NAST is the strongest prior method; nevertheless, UTRo-NAST-T outperforms our CTC-NAST* implementation on seven out of eight directions, with a small drop on Romanian (24.1 vs 24.7). Moreover, UTRo-NAST-C consistently improves over CTC-NAST* across all directions and increases the average BLEU from 28.2–29.2, highlighting the benefit of Conformer-based modeling for capturing acoustic cues and longer-range semantic dependencies.

Table 3. BLEU Scores and decoding speeds of various speech translation models evaluated on the MuST-C1st-COMMON set. The “pretrained model” columns indicate whether additional pretrained speech or text foundation models are used. Models marked with “*” denote our implementations

Model	Pretrained model		Text	De	Es	Fr	BLEU (↑)				Ro	Ru	Avg.	Speedup	Speedup*
	Speech						It	Nl	Pt						
Machine translation															
Transformer*															
Autoregressive models															
ESPhet-ST (Inaguma et al., 2020)					31.2	36.0	43.4	32.1	36.2	37.7	30.2	19.9	33.3	—	—
Fairseq S2T (Wang et al., 2020b)					22.9	28.0	32.7	23.8	27.4	28.0	21.9	15.8	25.1	—	—
FCCL (medium) (Zhang et al., 2023b)	✓				22.7	27.2	32.9	22.7	27.3	28.1	21.9	15.3	24.8	—	—
CMOT (Zhou et al., 2023)	✓				25.9	30.7	36.8	26.4	30.5	31.8	25.0	17.6	28.1	—	—
CTC-Aug ST (Xu et al., 2023)					27.0	31.1	37.3	26.9	31.2	32.7	25.3	17.9	28.7	—	—
CTC-Aug ST*					26.9	31.5	38.1	27.4	31.9	33.4	25.8	18.7	29.2	1.0x	—
CTC-Aug ST + Seq-KD (Xu et al., 2023)					27.0	32.6	38.5	27.9	31.8	32.9	25.3	18.0	29.3	—	1.0x
BiL-CTC + rescoring (Xu et al., 2024)					27.7	31.6	39.5	27.5	32.3	33.7	26.6	18.7	29.7	1.0x	—
Hard Multi-Task CTC/attn (Yan et al., 2024)	✓				28.4	32.0	39.5	28.2	33.0	34.3	27.0	19.6	30.3	—	—
ZEROSWOT-MEDIUM (Tsiamas et al., 2024a)	✓		✓		29.2	33.2	39.2	—	—	—	—	—	—	—	—
DiG-SST (Chen et al., 2024b)	✓		✓		30.5	34.9	39.4	30.6	35.0	37.1	27.8	20.3	31.9	—	—
ASR + AST SALM (Chen et al., 2024d)	✓		✓		26.9	30.9	37.6	—	—	—	—	—	—	—	—
GenTranslate (E2E) (Hu et al., 2024)	✓		✓		30.7	—	—	—	—	—	—	—	—	—	—
RoPE-ST+joint (Li et al., 2024)					—	33.9	—	29.4	—	—	—	—	—	—	—
S-Align (Zhang et al., 2024)	✓				23.9	28.4	34.2	23.9	28.4	29.5	22.9	16.2	25.9	—	—
Memory-ST (Yuan et al., 2024)					26.5	31.3	37.6	—	—	—	—	—	—	—	—
ST-MKD-MT (Wang et al., 2024)	✓				23.2	28.6	33.5	23.9	27.6	28.7	—	—	—	—	—
TDDR (Zhou et al., 2025)	✓				28.2	—	39.5	—	32.1	34.1	—	—	—	—	—
CAST (Tian et al., 2025)	✓		✓		29.8	35.3	41.6	28.5	—	—	—	—	—	—	—
PIA-8 (Xu et al., 2025)	✓				27.0	32.0	37.6	27.4	31.5	33.1	25.6	18.2	29.1	—	—
Non-autoregressive models															
CTC (Inaguma et al., 2021a)					24.1	29.0	34.6	24.3	28.5	—	—	—	—	13.83x	—
CMLM (Inaguma et al., 2021a)	✓				23.5	27.5	33.0	22.7	27.6	—	—	—	—	2.89x	—
Orthros-CTC (Inaguma et al., 2021b)					25.3	30.4	36.2	25.4	29.9	—	—	—	—	1.14x	—
Orthros-CMLM (Inaguma et al., 2021b)	✓				24.5	29.1	34.6	24.0	28.5	—	—	—	—	2.73x	—

(continued)

Table 3. Continued

Model	Pretrained model	Text	De	Es	Fr	It	BLEU (↑)				Ru	Avg.	Speedup	Speedup*
							Nl	Pt	Ro					
CTC-NAST (Xu et al., 2023)			25.8	—	—	—	—	—	—	—	—	—	5.67x	—
CTC-NAST*			25.9	31.9	37.5	26.5	30.6	31.8	24.7	16.9	28.2	—	—	4.55x
CTC-NAST + Seq-KD (Xu et al., 2023)			27.3	31.8	38.9	27.7	32.3	33.3	26.1	18.9	29.5	—	5.67x	—
Hard MultiTask CTC (Yan et al., 2024)	✓		23.4	28.4	33.6	—	—	—	—	—	—	—	—	—
NAST-S2T (Ma et al., 2024)			24.5	28.2	—	—	—	—	—	—	—	—	—	—
UTRo-NAST-T			26.7	32.0	37.8	27.5	31.2	32.0	24.1	17.5	28.6	—	—	4.74x
UTRo-NAST-T + Seq-KD			27.7	31.9	38.9	27.7	32.4	33.0	26.3	19.0	29.6	—	—	4.74x
UTRo-NAST-C			26.6	32.7	38.4	27.4	32.2	32.7	25.1	18.6	29.2	—	—	3.97x
UTRo-NAST-C + Seq-KD			27.3	32.1	38.7	28.0	32.5	33.5	26.3	18.9	29.7	—	—	3.97x

We also observe that applying sequence-level knowledge distillation (Seq-KD) tends to compress performance gaps among student architectures by aligning them to the teacher's output distribution, which explains why several Seq-KD variants fall into a narrow BLEU range (e.g. ~ 29.5 – 29.7) in Table 3. Under Seq-KD, the average BLEU differences become small: CTC-NAST + Seq-KD achieves 29.5, while UTRo-NAST-T + Seq-KD reaches 29.6 and UTRo-NAST-C + Seq-KD reaches 29.7. Importantly, the advantage of UTRo-NAST is more visible without Seq-KD, where architectural differences are less masked by distilled supervision (e.g. 28.2 for CTC-NAST* vs 28.6 for UTRo-NAST-T and 29.2 for UTRo-NAST-C). Consistent with this observation, the per-language-pair results suggest that UTRo-NAST yields robust gains across multiple directions, complementing the averaged BLEU comparison.

Beyond BLEU, UTRo-NAST offers practical advantages that are not directly reflected by aggregate scores. By explicitly separating monotonic word-level mapping from target-side reordering, UTRo-NAST provides an interpretable intermediate representation and enables targeted error diagnosis (e.g. distinguishing lexical mapping errors from reordering errors), which is less accessible in single-stage CTC baselines. In summary, UTRo-NAST provides consistent improvements over existing NAR-ST models without iterative refinement or complex decoding, while offering strong scalability, cross-lingual robustness and improved interpretability as a fully end-to-end ST framework.

5.3 Comparison with recent autoregressive-speech translation models

We also compare UTRo-NAST variants with recent autoregressive ST (AR-ST) systems, shown in the upper section of Table 3. AR-ST approaches such as CTC-Aug ST (Xu *et al.*, 2023), BiL-CTC + Rescoring (Xu *et al.*, 2024), ZEROSWOT-MEDIUM (Tsiamas *et al.*, 2024a) and CAST (Tian *et al.*, 2025) generally achieve strong performance, with average BLEU scores between 29 and 31. However, performance varies considerably across AR-ST models, from 25 to 31 BLEU. Many of these systems incorporate rescoring or rely on large pretrained foundation models for enhancement. In contrast, UTRo-NAST is trained in a purely end-to-end fashion without such external augmentation. Despite this, it surpasses several AR baselines, including FCCL (Medium) (Zhang *et al.*, 2023b), CMOT (Zhou *et al.*, 2023) and RoPE-ST+Joint (Li *et al.*, 2024), even without Seq-KD. With Seq-KD, UTRo-NAST achieves further gains and reaches performance comparable to stronger AR-ST baselines. Nonetheless, compared with top-performing AR models such as ZEROSWOT-MEDIUM, a gap remains, indicating that NAR-ST systems still trail AR models in ultimate accuracy.

5.4 Inference speed analysis

To assess efficiency, we measure the inference speed of UTRo-NAST against both AR-ST and NAR-ST baselines. Speed is evaluated on the En–De subset using an NVIDIA RTX 3090 GPU with a batch size of 1. As shown in Table 3, UTRo-NAST achieves a speedup of over $3.97\times$ relative to AR-ST models such as CTC-Aug ST. While UTRo-NAST-C typically produces better translations than UTRo-NAST-T, it incurs slightly higher latency due to the heavier computation of Conformer blocks. This illustrates a tradeoff between accuracy and efficiency. Overall, UTRo-NAST provides a strong balance: it consistently decodes faster than AR models while achieving comparable quality across multiple language pairs, demonstrating its practicality for real-time ST.

5.5 Ablation study on the word-by-word mapping encoder

To better isolate the contribution of the word-by-word mapping stage in our modular decomposition, we conduct a controlled ablation by setting the loss weight of the word-by-word mapping objective $\mathcal{L}_{\text{wbw}}$ to zero while keeping the network architecture unchanged. This setting removes the explicit word-level intermediate supervision but preserves the full forward computation of the word-by-word mapping encoder and the downstream reordering decoder, enabling a like-for-like comparison of translation quality.

Table 4 reports the BLEU scores on MuST-C `tst-COMMON`. For UTRo-NAST-T, disabling $\mathcal{L}_{\text{wbw}}$ by setting its weight to zero decreases the average BLEU from 28.6–28.4, with consistent drops on most language pairs (e.g. De: 26.7→25.9; Es: 32.0→31.2; Fr: 37.8→37.5; It: 27.5→27.3; Nl: 31.2→30.8). For UTRo-NAST-C, the degradation is more pronounced, where the average BLEU drops from 29.2–28.3, and decreases are observed across all directions (e.g. Fr: 38.4→37.1; Nl: 32.2→31.1; Pt: 32.7→31.2; Ro: 25.1→24.4). These results demonstrate that the explicit word-level intermediate supervision provided by $\mathcal{L}_{\text{wbw}}$ contributes meaningfully to the final translation quality, and that the benefit is particularly salient when the backbone has stronger sequence modeling capacity (i.e. the Conformer-based variant).

Overall, this ablation substantiates the importance of the proposed intermediate word-by-word mapping objective: even though the final translation is produced by the reordering decoder, enforcing monotonic word-level correspondence during training provides a useful inductive bias and improves end-to-end speech translation quality.

5.6 Understanding, translation and reordering for non-autoregressive speech translation with large language model-augmented correction

Finally, we evaluate the proposed inference-time correction framework that integrates LLMs, and compare it systematically with existing LLM-based ST approaches. In this study, we adopt the ChatGPT-4o API to implement the strategy, though the choice of LLM remains open depending on resources and deployment considerations. As shown in Table 5, we report results for UTRo-NAST-T and UTRo-NAST-C under different prompting strategies.

Several interesting observations can be made. First, both UTRo-NAST-T and UTRo-NAST-C degrade in performance when combined with the text-only prompt (TP). In contrast, ASR-Enhanced Prompting (AP) and ASR + Chain-of-Thought Prompting (ACoT) yield substantial improvements over the uncorrected baselines, confirming that LLMs can effectively leverage contextual information to enhance translation quality. These results signal the importance of ASR guidance: without it, corrections may drift from the original content, thereby reducing translation fidelity.

Table 4. Ablation results on the MuST-C `tst-COMMON` set, comparing the full UTRo-NAST models with an ablated setting where the word-by-word mapping loss $\mathcal{L}_{\text{wbw}}$ is disabled (loss weight set to zero) while keeping the architecture unchanged

Model	BLEU (↑)								Avg.
	De	Es	Fr	It	Nl	Pt	Ro	Ru	
UTRo-NAST-T	26.7	32.0	37.8	27.5	31.2	32.0	24.1	17.5	28.6
– $\mathcal{L}_{\text{wbw}}$	25.9	31.2	37.5	27.3	30.8	32.3	24.5	17.7	28.4
UTRo-NAST-C	26.6	32.7	38.4	27.4	32.2	32.7	25.1	18.6	29.2
– $\mathcal{L}_{\text{wbw}}$	26.2	31.7	37.1	26.7	31.1	31.2	24.4	18.0	28.3

Table 5. Performance of UTRo-NAST paired with LLM-augmented correction on MuST-C test - COMMON

Model	BLEU (↑)					Pt	Ro	Ru	Avg.
	De	Es	Fr	It	Nl				
UTRo-NAST-T	26.68	32.03	37.84	27.45	31.24	31.95	24.09	17.48	28.60
+LLM-Augmented correction									
Text-Only prompt (TP)	26.62	32.08	33.74	27.50	30.94	28.35	22.69	18.51	27.55
ASR-Enhanced prompt (AP)	31.01	34.77	40.69	30.84	34.21	32.58	27.44	22.14	31.71
ASR + chain-of-thought prompt (ACoT)	30.28	33.68	40.13	29.62	32.84	31.16	26.64	21.65	30.75
UTRo-NAST-C	26.63	32.68	38.35	27.35	32.17	32.68	25.08	18.63	29.20
+LLM-Augmented correction									
Text-Only prompt (TP)	26.52	32.38	33.87	27.16	31.20	29.01	23.40	19.03	27.82
ASR-Enhanced prompt (AP)	30.87	34.73	40.88	30.94	34.13	32.65	27.44	22.30	31.74
ASR + chain-of-thought prompt (ACoT)	30.30	33.61	40.06	29.54	32.96	31.21	26.62	21.67	30.75
LST-7B (Zhang et al., 2023a)	29.15	33.07	40.7	—	—	—	—	—	—
LST-13B (Zhang et al., 2023a)	30.39	35.33	41.55	—	—	—	—	—	—
ASR + AST SALM (Chen et al., 2024e)	30.7	—	—	—	—	—	—	—	—
GenTranslate (Hu et al., 2024)	—	33.9	—	29.4	—	—	—	—	—
GenTranslate-V2 (Hu et al., 2024)	—	32.2	—	27.3	—	—	—	—	—
Refine _{path} -GPT3.5(R1) (Dou et al., 2025)	30.52	34.03	40.99	—	—	—	—	—	—
Refine _{path} -LLaMA3-8B (Dou et al., 2025)	32.43	35.62	42.93	—	—	—	—	—	—
Refine _{path} -mistral-12B (Dou et al., 2025)	33.39	36.08	44.22	—	—	—	—	—	—
TDR (Zhou et al., 2025)	29.8	35.3	41.6	28.5	—	—	—	—	—

Next, somewhat surprisingly, AP achieves better performance than ACoT. A plausible explanation is that while CoT encourages more detailed reasoning, it may also introduce unnecessary complexity or overcorrection, leading to minor semantic deviations. In contrast, AP provides a more direct and efficient alignment between source ASR output and translation, which appears more reliable in this setting.

Finally, we compare LLM-augmented correction with recent state-of-the-art LLM-based ST systems, including GenTranslate, Refine_{Both}, and TDR. Even without fine-tuning the LLM, our modular framework with carefully designed prompts achieves results comparable to or exceeding these systems. For example, on German (De), UTRo-NAST-T with AP reaches 31.01 BLEU, outperforming Refine_{Both}-GPT3.5 (30.52) and TDR (29.8). On Italian (It), it achieves 30.84 BLEU, surpassing GenTranslate (29.4). These findings highlight the strong cross-lingual consistency and generalization ability of the proposed framework.

5.7 Discussion on the large language model-augmented correction strategy

5.7.1 Latency and cost considerations. As shown in Table 5, incorporating the LLM correction step (AP or ACoT) yields clear improvements over the uncorrected baseline in our main setting (ChatGPT-4o). However, these gains come with additional inference latency and monetary cost, especially when the correction module is deployed via a commercial API.

To provide rough estimates for real-world usage, we measure the latency of the LLM-augmented correction on the entire En-De test-COMMON set (average sentence length ≈ 17 words). Using the ChatGPT-4o API in our setting, the average per-sentence latency is approximately 1.60 s for AP and 1.70 s for ACoT. These numbers reflect an API-based deployment and may vary with network conditions and service load; nevertheless, they provide a concrete reference for the incremental overhead introduced by enabling postcorrection.

Regarding cost, it scales approximately linearly with the number of input/output tokens under token-based pricing. Therefore, in practical deployments, one should weigh the translation-quality gains against the additional latency and API expenses, particularly under real-time constraints.

5.7.2 Case study. We present two qualitative case studies (En→De) to illustrate both the benefits and potential failure modes of the proposed LLM-augmented correction strategy. As

Table 6. Qualitative example (En → De) where TP exhibits semantic drift (e.g. incorrect possessive “seines lebens” and gendered noun choice “hausbesitzer” vs “hausbesitzerin”), while AP/ACoT recover the intended meaning via ASR grounding and structured refinement

Item	Text
SRC	When I found the captain, he was having a very engaging conversation with the homeowner, who was surely having one of the worst days of her life
REF	Als ich den Hauptmann fand, hatte er gerade in eine sehr ernste Unterhaltung mit der Hausbesitzerin, die sicherlich einen der schlimmsten Tage ihres Lebens hatte
UTRo-NAST-T	Als ich den Kapitän fand, hatte er ein sehr einnehmendes Gespräch mit dem Hausbesitzer, der sicherlich eines der schlimmsten Tage ihres Lebens hatte
+TP	Als ich den Kapitän fand, führte er ein sehr einnehmendes Gespräch mit dem Hausbesitzer, der sicherlich einen der schlimmsten Tage seines Lebens erlebte
+AP	Als ich den Kapitän fand, führte er ein sehr fesselndes Gespräch mit der Hausbesitzerin, die sicherlich einen der schlimmsten Tage ihres Lebens erlebte
+ACoT	Als ich den Kapitän fand, führte er ein sehr fesselndes Gespräch mit der Hausbesitzerin, die sicherlich einen der schlimmsten Tage ihres Lebens hatte

shown in Table 6, text-only prompting (TP) may occasionally introduce semantic drift by altering core attributes such as possessives or gendered noun choices (e.g. “seines Lebens” and “Hausbesitzer” vs “Hausbesitzerin”), leading to a mismatch with the reference. In contrast, AP/ACoT incorporate ASR grounding (i.e. the transcript context) and structured refinement, which helps preserve the intended meaning and recovers a translation closer to the reference in this example.

Table 7 further highlights that the correction behavior is not always monotonic. While TP improves the baseline output by reducing disfluency and repetition, AP/ACoT may over-edit the hypothesis by introducing redundant phrasing (e.g. “und das sind diejenigen”) or heavier rephrasing such as expanded relative-clause rewrites (e.g. “aus denen ich das meiste Feedback erhalten habe”). These cases indicate that, while LLM-based postcorrection generally improves translation quality, it is not flawless. In particular, the correction model may occasionally hallucinate or over-generate content (i.e. unnecessary expansions or rephrasings) that deviates from the reference. Addressing such non-monotonic behaviors remains an important direction for future work toward more robust and reliable LLM-based refinement.

5.7.3 Cross-large language model robustness of prompting variants. To assess whether the observed behaviors of different prompting variants are robust across LLM backends (rather than being specific to a single model), we evaluate the proposed LLM-augmented correction using two LLMs: ChatGPT-4o and a smaller open-weight model, Mistral (Bérard *et al.*, 2016), which has 12B parameters. The results are summarized in Table 8. Across both backends and for both UTRo-NAST-T and UTRo-NAST-C, we observe a consistent ordering among prompting variants: AP achieves higher BLEU than ACoT, and both outperform TP. These results suggest that the AP-over-ACoT trend is robust across LLM choices and is not limited to a particular proprietary model.

Meanwhile, the absolute impact of postcorrection is clearly LLM-dependent. While ChatGPT-4o provides substantial gains with AP/ACoT, Mistral-12B does not surpass the

Table 7. Qualitative example (En → De) where TP improves the baseline output, but AP/ACoT can degrade performance due to potential over-editing: AP introduces redundant phrasing (e.g. “und das Sind diejenigen”), while ACoT performs an expanded relative-clause rewrite (e.g. “aus denen ich das meiste feedback erhalten habe”) that may deviate from the reference despite remaining grammatical

Item	Text
SRC	What’s incredible for me is the projects that I’ve gotten the most feedback out of, or I’ve had the most success in, are ones where I’ve interacted with things directly.
REF	Das Unglaubliche für mich ist, dass die Projekte von denen ich das meiste Feedback bekam oder wo ich den größten Erfolg hatte, jene sind, wo ich mit Dingen direkt interagiere
UTRo-NAST-T	Was unglaublich für mich ist, dass die Projekte, die ich bekommen habe, das meiste Feedback habe, die ich am meisten Erfolg hatte, und diejenigen sind, in denen ich mit Dingen direkt interagiert habe
+TP	Was für mich unglaublich ist, ist, dass die Projekte, bei denen ich das meiste Feedback und den größten Erfolg hatte, diejenigen sind, bei denen ich direkt mit den Dingen interagiert habe
+AP	Was für mich unglaublich ist, sind die Projekte, bei denen ich das meiste Feedback und den größten Erfolg hatte, und das sind diejenigen, bei denen ich direkt mit den Dingen interagiert habe
+ACoT	Was für mich unglaublich ist, ist, dass die Projekte, aus denen ich das meiste Feedback erhalten habe, und am erfolgreichsten war, diejenigen sind, bei denen ich direkt mit den Dingen interagiert habe

Table 8. BLEU Scores of the LLM-Augmented correction strategy with different LLM backends using the same prompting variants. The uncorrected baselines are included for reference. Mistral-12B denotes the mistral API model open-mistral-nemo-2407 (mistral NeMo 12B, v24.07)

LLM	UTRo-NAST-T				UTRo-NAST-C			
	w/o correction	TP	AP	ACoT	w/o correction	TP	AP	ACoT
ChatGPT-4o		27.55	31.71	30.75		27.82	31.74	30.75
Mistral-12B	28.60	24.32	27.13	26.10	29.20	24.74	27.18	26.06

uncorrected baseline in our current setting and may even degrade BLEU. A plausible explanation is that, as a smaller model, Mistral-12B may have weaker instruction-following or refinement capability for this postediting style task, making it more prone to suboptimal edits. These results highlight an important practical consideration: the proposed prompting framework appears transferable across LLMs in terms of relative behaviors, but the magnitude (and even the sign) of the quality change depends on the underlying LLM capability. Improving the robustness of LLM-based refinement for smaller or open-weight LLMs remains an important direction for future work.

6. Conclusion

In this work, we proposed UTRo-NAST, a non-autoregressive speech translation framework that decomposes the complex ST process into three modular subtasks: source speech understanding, word-level translation, and target-side reordering. By adopting a divide-and-conquer philosophy with multitask optimization, UTRo-NAST achieves competitive translation quality while maintaining high inference efficiency and architectural flexibility. Experiments on eight language pairs show that UTRo-NAST consistently outperforms existing NAR-ST baselines and achieves performance comparable to strong autoregressive models.

Beyond the core framework, we further introduced a plug-and-play LLM-augmented correction strategy that refines preliminary translations through prompting-based correction. This lightweight extension significantly improves fluency and robustness, achieving results competitive with state-of-the-art LLM-based ST systems, while avoiding iterative refinement or costly LLM fine-tuning.

Overall, these findings show the effectiveness and scalability of UTRo-NAST for real-world speech translation. For future work, we plan to explore its application in low-resource and real-time scenarios, investigate tighter integration of large language models to further boost translation quality, and extend the framework to multimodal translation tasks.

Acknowledgements

This work was supported by the National Science and Technology Council (NSTC), Taiwan, under Grants NSTC 112-2628-E-011-008-MY3 and NSTC 114-2640-B-002-005. Additional support was provided by the Empower Vocational Education Research Center at the National Taiwan University of Science and Technology (NTUST) through the Featured Areas Research Center Program, as part of the Higher Education Sprout Project funded by the Ministry of Education (MOE), Taiwan. The authors also acknowledge the National Center for High-Performance Computing, National Applied Research Laboratories (NARLabs), Taiwan, for providing essential computational and storage resources.

References

- Bérard, A., Pietquin, O., Besacier, L. and Servan, C. (2016), “Listen and translate: a proof of concept for end-to-end speech-to-text translation”, *NIPS Workshop on End-to-End Learning for Speech and Audio Processing*, Barcelona, Spain, available at: <https://hal.science/hal-01408086>
- Chen, X., Zhang, S., Bai, Q., Chen, K. and Nakamura, S. (2024a), “LLaST: improved end-to-end speech translation system leveraged by large language models”, Ku, L.W., Martins, A. and Srikumar, V. (Eds), *Findings of the Association for Computational Linguistics: ACL 2024*, Association for Computational Linguistics, Bangkok, Thailand, pp. 6976-6987, doi: [10.18653/v1/2024.findings-acl.416](https://doi.org/10.18653/v1/2024.findings-acl.416), available at: <https://aclanthology.org/2024.findingsacl.416/>
- Chen, X., Fan, K., Luo, W., Zhang, L., Zhao, L., Liu, X. and Huang, Z. (2024b), “Divergence-guided simultaneous speech translation”, *Proc. of AAAI*.
- Chen, X., Fan, K., Luo, W., Zhang, L., Zhao, L., Liu, X. and Huang, Z. (2024c), “Divergence-guided simultaneous speech translation”, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38 No. 16, pp. 17799-17807, doi: [10.1609/aaai.v38i16.29733](https://doi.org/10.1609/aaai.v38i16.29733), available at: <https://ojs.aaai.org/index.php/AAAI/article/view/29733>
- Chen, Z., Huang, H., Andrusenko, A., Hrinchuk, O., Puvvada, K.C., Li, J., Ghosh, S., Balam, J. and Ginsburg, B. (2024d), “Salm: speech augmented language model with in-context learning for speech recognition and translation”, *Proc. of ICASSP*.
- Chen, Z., Huang, H., Andrusenko, A., Hrinchuk, O., Puvvada, K.C., Li, J., Ghosh, S., Balam, J. and Ginsburg, B. (2024e), “SALM: speech-augmented language model with in-context learning for speech recognition and translation”, *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13521-13525, doi: [10.1109/ICASSP48485.2024.10447553](https://doi.org/10.1109/ICASSP48485.2024.10447553).
- Cheng, X., Dong, Q., Yue, F., Ko, T., Wang, M. and Zou, Y. (2023), “M3ST: mix at three levels for speech translation”, *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5, doi: [10.1109/ICASSP49357.2023.10095090](https://doi.org/10.1109/ICASSP49357.2023.10095090).
- Chuang, S.P., Chuang, Y.S., Chang, C.C. and Lee, H.Y. (2021), “Investigating the reordering capability in CTC-based non-autoregressive end-to-end speech translation”, in Zong, C., Xia, F., Li, W. and Navigli, R. (Eds), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online: Association for Computational Linguistics, pp. 1068-1077, doi: [10.18653/v1/2021.findings-acl.92](https://doi.org/10.18653/v1/2021.findings-acl.92), available at: <https://aclanthology.org/2021.findings-acl.92/>
- Di Gangi, M.A., Cattoni, R., Bentivogli, L., Negri, M. and Turchi, M. (2019), “MuST-C: a multilingual speech translation corpus”, in Burstein, J., Doran, C. and Solorio, T. (Eds), *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)*, Association for Computational Linguistics, Minneapolis, MN, pp. 2012-7, doi: [10.18653/v1/N19-1202](https://doi.org/10.18653/v1/N19-1202), available at: <https://aclanthology.org/N19-1202/>
- Dou, H., Tian, X., Lyu, X., Zhu, J., Li, J. and Guo, L. (2025), “Speech translation refinement using large language models”, *arXiv preprint arXiv:2501.15090*.
- Fang, Q. and Feng, Y. (2023), “Understanding and bridging the modality gap for speech translation”, in Rogers, A., Boyd-Graber, J. and Okazaki, N. (Eds), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Toronto, Canada, pp. 15864-15881, doi: [10.18653/v1/2023.acl-long.884](https://doi.org/10.18653/v1/2023.acl-long.884), available at: <https://aclanthology.org/2023.acl-long.884/>
- Fang, Q., Ye, R., Li, L., Feng, Y. and Wang, M. (2022), “STEMM: self-learning with speech-text manifold mix-up for speech translation”, in Muresan, S., Nakov, P. and Villavicencio, A. (Eds), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Dublin, Ireland, pp. 7050-7062, doi: [10.18653/v1/2022.acl-long.486](https://doi.org/10.18653/v1/2022.acl-long.486), available at: <https://aclanthology.org/2022.acl-long.486/>

- Gao, P., Zhang, R., He, Z., Wu, H. and Wang, H. (2024), "An empirical study of consistency regularization for end-to-end speech-to text translation", in Duh, K., Gomez, H. and Bethard, S. (Eds), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Association for Computational Linguistics, Mexico City, Mexico, pp. 242-256, doi: [10.18653/v1/2024.naacl-long.14](https://doi.org/10.18653/v1/2024.naacl-long.14), available at: <https://aclanthology.org/2024.naacl-long.14/>
- Ghazvininejad, M., Levy, O., Liu, Y. and Zettlemoyer, L. (2019), "Mask-predict: parallel decoding of conditional masked language models", in Inui, K., Jiang, J., Ng, V. and Wan, X. (Eds), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, pp. 6112-6121, doi: [10.18653/v1/D19-1633](https://doi.org/10.18653/v1/D19-1633), available at: <https://aclanthology.org/D19-1633/>
- Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J. (2006), "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks", *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*, Association for Computing Machinery, Pittsburgh, PA, pp. 369-376, isbn: 1595933832, doi: [10.1145/1143844.1143891](https://doi.org/10.1145/1143844.1143891).
- Gu, J., Bradbury, J., Xiong, C., Li, V.O. and Socher, R. (2018), "Non-autoregressive neural machine translation", *International Conference on Learning Representations*, available at: <https://openreview.net/forum?id=B1l8BtlCb>
- Gulati, A., Qin, J., Chiu, C.C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al (2020), "Conformer: convolution augmented transformer for speech recognition", *arXiv preprint arXiv:2005.08100*.
- Hu, Y., Chen, C., Yang, C.H.H., Li, R., Zhang, D., Chen, Z. and Chng, E.S. (2024), "GenTranslate: large language models are generative multilingual speech and machine translators", in Ku, L.W., Martins, A. and Srikumar, V. (Eds), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Bangkok, Thailand, pp. 74-90, doi: [10.18653/v1/2024.acl-long.5](https://doi.org/10.18653/v1/2024.acl-long.5), available at: <https://aclanthology.org/2024.acl-long.5/>
- Hu, K., Chen, Z., Yang, C.H.H., Żelasko, P., Hrinchuk, O., Lavrukhin, V., Balam, J. and Ginsburg, B. (2025), "Chain-of-thought prompting for speech translation", *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5, doi: [10.1109/ICASSP49660.2025.10890560](https://doi.org/10.1109/ICASSP49660.2025.10890560).
- Inaguma, H., Higuchi, Y., Duh, K., Kawahara, T. and Watanabe, S. (2021a), "Non-autoregressive end-to-end speech translation with parallel autoregressive rescoring", *arXiv preprint arXiv:2109.04411*.
- Inaguma, H., Higuchi, Y., Duh, K., Kawahara, T. and Watanabe, S. (2021b), "ORTHROS: non-autoregressive end-to-end speech translation with dual-decoder", *ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7503-7507, doi: [10.1109/ICASSP39728.2021.9415093](https://doi.org/10.1109/ICASSP39728.2021.9415093).
- Inaguma, H., Kiyono, S., Duh, K., Karita, S., Yalta, N., Hayashi, T. and Watanabe, S. (2020), "ESPnet-ST: all-in-one speech translation toolkit", in Celikyilmaz, A. and Wen, T.H. (Eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, pp. 302-311, doi: [10.18653/v1/2020.acl-demos.34](https://doi.org/10.18653/v1/2020.acl-demos.34), available at: <https://aclanthology.org/2020.acl-demos.34/>
- Kim, Y. and Rush, A.M. (2016), "Sequence-level knowledge distillation", *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Su, J., Duh, K. and Carreras, X. (Eds), Association for Computational Linguistics, Austin, TX, pp. 1317-1327, doi: [10.18653/v1/D16-1139](https://doi.org/10.18653/v1/D16-1139), available at: <https://aclanthology.org/D16-1139/>
- Kuan, Y.C. and Chen, K.Y. (2025), "UTRo-NAST: non-autoregressive speech translation via understanding, translation, and reordering", *Proc. of APSIPA*.

- Lee, J. and Watanabe, S. (2021), “Intermediate loss regularization for CTC-based speech recognition”, *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6224-6228, doi: [10.1109/ICASSP39728.2021.9414594](https://doi.org/10.1109/ICASSP39728.2021.9414594).
- Li, X., Li, S., Zhang, X.L. and Rahardja, S. (2024), “Transformer-based end-to-end speech translation with rotary position embedding”, *IEEE Signal Processing Letters*, Vol. 31, pp. 371-375.
- Ma, Z., Fang, Q., Zhang, S., Guo, S., Feng, Y. and Zhang, M. (2024), “A non-autoregressive generation framework for end-to-end simultaneous speech-to-any translation”, *Proc. of ACL*.
- Nozaki, J. and Komatsu, T. (2021), “Relaxing the conditional independence assumption of CTC-based ASR by conditioning on intermediate predictions”, arXiv: 2104.02724 [eess.AS], available at: <https://arxiv.org/abs/2104.02724>
- Omachi, M., Yan, B., Dalmia, S., Fujita, Y. and Watanabe, S. (2023), “Align, write, re-order: explainable end-to-end speech translation via operation sequence generation”, *Proc. of ICASSP*.
- Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D. and Le, Q.V. (2019), “SpecAugment: a simple data augmentation method for automatic speech recognition”, *Interspeech*, available at: <https://api.semanticscholar.org/CorpusID:121321299>
- Post, M. (2018), “A call for clarity in reporting BLEU scores”, in Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A.J., Koehn, P., Monz, C., Negri, M., N  v  l, A., Neves, M., Post, M., Specia, L., Turchi, M. and Verspoor, K. (Eds), *Proceedings of the Third Conference on Machine Translation: Research Papers, Association for Computational Linguistics*, Brussels, Belgium, pp. 186-191, doi: [10.18653/v1/W18-6319](https://doi.org/10.18653/v1/W18-6319), available at: <https://aclanthology.org/W18-6319/>
- Qian, L., Zhou, H., Bao, Y., Wang, M., Qiu, L., Zhang, W., Yu, Y. and Li, L. (2021), “Glancing transformer for non-autoregressive neural machine translation”, in Zong, C., Xia, F., Li, W. and Navigli, R. (Eds), *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)*, Online: Association for Computational Linguistics, pp. 1993-2003, doi: [10.18653/v1/2021.acl-long.155](https://doi.org/10.18653/v1/2021.acl-long.155), available at: <https://aclanthology.org/2021.acl-long.155/>
- Ran, Q., Lin, Y., Li, P. and Zhou, J. (2021), “Guiding non-autoregressive neural machine translation decoding with reordering information”, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35 No. 15, pp. 13727-13735, doi: [10.1609/aaai.v35i15.17618](https://doi.org/10.1609/aaai.v35i15.17618), available at: <https://ojs.aaai.org/index.php/AAAI/article/view/17618>
- Stentiford, F.W.M. and Steer, M.G. (1990), “Machine translation of speech”, *Speech and Language Processing*, Chapman and Hall, Ltd., GBR, pp. 183-196. isbn: 0442312075.
- Tian, X., Wei, H., Gong, Z., Li, J. and Xie, J. (2025), “Improving end-to-end speech-to-text translation with document-level context”, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 33, pp. 2098-2109, doi: [10.1109/TASLPRO.2025.3570951](https://doi.org/10.1109/TASLPRO.2025.3570951). issn2998-4173
- Tsiamas, I., G  llego, G., Fonollosa, J. and Costa-Juss  , M. (2024a), “Pushing the limits of zero-shot end-to-end speech translation”, *Proc. of ACL*.
- Tsiamas, I., G  llego, G.I., Fonollosa, J.A.R. and Costajuss  , M.R. (2024b), “Pushing the limits of zero-shot end-to-end speech translation”, in Ku, L.W., Martins, A. and Srikumar, V. (Eds), *Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics*, Bangkok, Thailand, pp. 14245-14267, doi: [10.18653/v1/2024.findings-acl.847](https://doi.org/10.18653/v1/2024.findings-acl.847), available at: <https://aclanthology.org/2024.findings-acl.847/>
- Wang, C., Wu, Y., Liu, S., Zhou, M. and Yang, Z. (2020a), “Curriculum pre-training for end-to-end speech translation”, in Jurafsky, D., Chai, J., Schluter, N. and Tetreault, J. (Eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics*, pp. 3728-3738, doi: [10.18653/v1/2020.acl-main.344](https://doi.org/10.18653/v1/2020.acl-main.344), available at: <https://aclanthology.org/2020.acl-main.344/>

- Wang, C., Tang, Y., Ma, X., Wu, A., Okhonko, D. and Pino, J. (2020b), "Fairseq S2T: fast speech-to-text modeling with fairseq", in Wong, D. and Kiela, D. (Eds), *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: System Demonstrations, Association for Computational Linguistics*, Suzhou, China, pp. 33-39, doi: [10.18653/v1/2020.aacl-demo.6](https://doi.org/10.18653/v1/2020.aacl-demo.6), available at: <https://aclanthology.org/2020.aacl-demo.6/>
- Wang, H., Xue, Z., Lei, Y. and Xiong, D. (2024), "End-to-end speech translation with mutual knowledge distillation", *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11306-11310, doi: [10.1109/ICASSP48485.2024.10445811](https://doi.org/10.1109/ICASSP48485.2024.10445811).
- Weiss, R.J., Chorowski, J., Jaitly, N., Wu, Y. and Chen, Z. (2017a), "Sequenceto-sequence models can directly translate foreign speech", *Interspeech*, available at: <https://api.semanticscholar.org/CorpusID:7857444>
- Weiss, R.J., Chorowski, J., Jaitly, N., Wu, Y. and Chen, Z. (2017b), "Sequenceto-sequence models can directly translate foreign speech", *Interspeech 2017*, pp. 2625-2629, doi: [10.21437/Interspeech.2017-503](https://doi.org/10.21437/Interspeech.2017-503).
- Xu, C., Liu, X., Zhang, Y., Ma, A., Xiao, T., Zhu, J., Man, D. and Yang, W. (2025), "Unveiling the fundamental obstacle in speech-to-text modeling: understanding and mitigating the granularity challenge", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 33, pp. 1719-1729.
- Xu, C., Liu, X., He, E., Zhang, Y., Dong, Q., Xiao, T., Zhu, J., Man, D. and Yang, W. (2024), "Bridging the gaps of both modality and language: synchronous bilingual CTC for speech translation and speech recognition", *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12176-12180, doi: [10.1109/ICASSP48485.2024.10445849](https://doi.org/10.1109/ICASSP48485.2024.10445849).
- Xu, C., Liu, X., Liu, X., Sun, Q., Zhang, Y., Yang, M., Dong, Q., Ko, T., Wang, M., Xiao, T., Ma, A. and Zhu, J. (2023), "CTC-based nonautoregressive speech translation", in Rogers, A., Boyd-Graber, J. and Okazaki, N. (Eds), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Toronto, Canada, pp. 13321-13339, doi: [10.18653/v1/2023.acl-long.744](https://doi.org/10.18653/v1/2023.acl-long.744), available at: <https://aclanthology.org/2023.acl-long.744/>
- Yan, B., Chang, X., Anastasopoulos, A., Fujita, Y. and Watanabe, S. (2024), "Cross-modal multi-tasking for speech-to-text translation via hard parameter sharing", *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11941-11945, doi: [10.1109/ICASSP48485.2024.10447926](https://doi.org/10.1109/ICASSP48485.2024.10447926).
- Yan, B., Dalmia, S., Higuchi, Y., Neubig, G., Metze, F., Black, A.W. and Watanabe, S. (2023), "CTC alignments improve autoregressive translation", in Vlachos, A. and Augenstein, I. (Eds), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics*, Dubrovnik, Croatia, pp. 1623-1639, doi: [10.18653/v1/2023.eacl-main.119](https://doi.org/10.18653/v1/2023.eacl-main.119), available at: <https://aclanthology.org/2023.eacl-main.119/>
- Ye, R., Wang, M. and Li, L. (2021), "End-to-end speech translation via cross-modal progressive training", *Interspeech 2021*, pp. 2267-2271, doi: [10.21437/Interspeech.2021-1065](https://doi.org/10.21437/Interspeech.2021-1065).
- Ye, R., Wang, M. and Li, L. (2022), "Cross-modal contrastive learning for speech translation", in Carpuat, M., de Marneffe, M.C. and Meza Ruiz, I.V. (Eds), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics*, Seattle, United States, pp. 5099-5113, doi: [10.18653/v1/2022.naacl-main.376](https://doi.org/10.18653/v1/2022.naacl-main.376), available at: <https://aclanthology.org/2022.naacl-main.376/>
- Yuan, Y., Zhou, Y. and Shi, X. (2024), "Memory-augmented speech-to-text translation with multi-scale context translation strategy", *Proc. of ICASSP*.

- Zhang, H., Si, N., Chen, Y., Zhang, W., Yang, X., Qu, D. and Jiao, X. (2023a), "Tuning large language model for end-to-end speech translation", CoRR, abs/2310.02050, available at: 10.48550/arXiv.2310.02050.
- Zhang, H., Si, N., Chen, Y., Zhang, W., Yang, X., Qu, D. and Zhang, W.Q. (2023b), "Improving speech translation by cross-modal multi-grained contrastive learning", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 31, pp. 1075-1086, doi: [10.1109/TASLP.2023.3244521](https://doi.org/10.1109/TASLP.2023.3244521).)
- Zhang, Y., Kou, K., Li, B., Xu, C., Zhang, C., Xiao, T. and Zhu, J. (2024), "Soft alignment of modality space for end-to-end speech translation", *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11041-11045, doi: [10.1109/ICASSP48485.2024.10447494](https://doi.org/10.1109/ICASSP48485.2024.10447494).
- Zhao, C., Wang, M., Dong, Q., Ye, R. and Li, L. (2021), "NeurST: neural speech translation toolkit", in Ji, H., Park, J.C. and Xia, R. (Eds), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, Online: Association for Computational Linguistics*, pp. 55-62, doi: [10.18653/v1/2021.acl-demo.7](https://doi.org/10.18653/v1/2021.acl-demo.7), available at: <https://aclanthology.org/2021.acl-demo.7/>
- Zhou, Y., Fang, Q. and Feng, Y. (2023), "CMOT: cross-modal mixup via optimal transport for speech translation", in Rogers, A., Boyd-Graber, J. and Okazaki, N. (Eds), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Toronto, Canada, pp. 7873-7887, doi: [10.18653/v1/2023.acl-long.436](https://doi.org/10.18653/v1/2023.acl-long.436), available at: <https://aclanthology.org/2023.acl-long.436/>
- Zhou, Y., Yuan, Y., Zhang, C. and Shi, X. (2025), "Boosting context-aware speech translation with large language models", *IEEE Signal Processing Letters*, Vol. 32, pp. 1955-1959.

Corresponding author

Kuan-Yu Chen can be contacted at: kychen@mail.ntust.edu.tw