PLaST: Towards Paralinguistic-aware Speech Translation

Yi Li[1,2,3*], **Rui Zhao**[1,2,3*], **Ruiquan Zhang**[1,2,3],
Jinsong Su[1,2,3], **Daimeng Wei**[4], **Min Zhang**[4], **Yidong Chen**[1,2,3†]

[1]Department of Artificial Intelligence, School of Informatics, Xiamen University, China
[2]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, China
[3]National Language Resources Monitoring and Research Center for Education and Teaching Media, Xiamen University, China
[4]Huawei Translation Services Center, Beijing, China
{liangyi, zhsqzr}@stu.xmu.edu.cn, ydchen@xmu.edu.cn

Abstract

Speech translation (ST) aims to translate speech from a source language into text in the target language. Naturally, speech signals contain paralinguistic cues beyond linguistic content, which could influence or even alter the interpretation of a lexically identical sentence, thereby yielding distinct translations. However, existing ST models lack direct and sufficient modeling of paralinguistic information, which limits their ability to perceive paralinguistic cues and understand speech comprehensively, leading to degraded translation performance. In response, we propose **P**ara**L**inguistic-**a**ware **S**peech **T**ranslation (**PLaST**), a novel dual-branch framework which directly leverages paralinguistic cues beyond the linguistic content. Specifically, PLaST employs a speech encoder and a style extractor to independently generate linguistic and paralinguistic representations, respectively. To obtain a purified linguistic representation aligned with the text representation, a hierarchical Optimal Transport (**OT**) is applied on the layer-wise outputs from an LLM decoder. Then, the paralinguistic information is retrieved and refined with an Attention-based Retrieval (**AR**) module, with the linguistic representation serving as queries to enable joint guidance for semantic understanding and translation generation. PLaST outperforms the strong baseline with an average of **5.0** directional and **4.5** global contrastive likelihood scores on the paralinguistic-sensitive benchmark ContraProST, demonstrating its superior capability in paralinguistic perception. Further experiments on the standard speech translation benchmark CoVoST-2 show that PLaST generalizes well to typical ST scenarios.

Code — https://github.com/YancyDan/PLaST

1 Introduction

Speech translation (ST) refers to the task of converting speech signals in one language into written text in another. In addition to *linguistic* content such as morphology, syntax, and semantics found in the literal source transcript, speech signals also carry *paralinguistic* cues, including emphasis, emotion, and prosody (Bolinger 1961; Banse and Scherer 1996; Shriberg et al. 1998), which play crucial roles in focusing and clarifying meaning, disambiguating syntax and sentence structure, conveying the emotional state of the speaker (Lin, Chiang, and Lee 2024; Tsiamas et al. 2024b). These factors contribute to inherent representation differences between text and speech, making ST a particularly challenging task. As illustrated by the two examples in Figure 1, while the linguistic components deliver the core message, variations in paralinguistic features can influence or even alter the intended meaning of an utterance (Castro et al. 2019), leading to disparate translations. Therefore, investigating the effective utilization of paralinguistic information represents a promising research direction for improving the

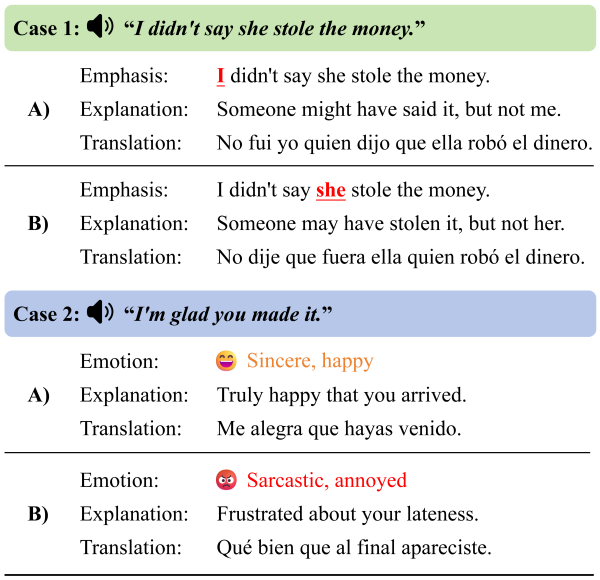


Figure 1: The same sentence can be interpreted differently when spoken with diverse paralinguistic cues, resulting in distinct translations from English to Spanish. Case 1: Influence of Emphasis. Case 2: Influence of Emotion.

*These authors contributed equally.
†Corresponding Author.

performance of speech translation.

Traditional ST first performs an automatic speech recognition (ASR) and then translates the transcribed transcript with a machine translation (MT) module (Sperber and Paulik 2020; Han et al. 2021). The cascaded approaches discard the original speech signals and rely solely on predicted transcripts containing only linguistic information during translation. In recent years, end-to-end speech translation (E2E ST) has become a mainstream paradigm as an alternative to cascaded ST by directly translating source speech into target text (Tsiamas et al. 2024a; Fang and Feng 2023; Zhou, Fang, and Feng 2023), which could be further enhanced by the impressive semantic understanding and language modeling capabilities of large language models (LLMs) (Chen et al. 2024b; Zhang et al. 2023a). As a typical cross-modal task, E2E ST suffers from the modality gap issues and various approaches have been proposed to align the representations of source speech and text, including multi-task learning (Tang et al. 2021; Ye, Wang, and Li 2021), cross-modal mix-up (Zhou, Fang, and Feng 2023; Fang et al. 2022), contrastive learning (Ouyang, Ye, and Li 2023; Ye, Wang, and Li 2022), and adversarial training (Zhang et al. 2024b). However, the focus on aligning source speech with corresponding textual content has led to a predominant reliance on linguistic information, which may limit the models' ability to exploit paralinguistic cues since it encourages models to treat speech signals primarily in terms of their linguistic content. Meanwhile, LLM-integrated methods typically employ relatively simple adaptation modules to align speech representations with the LLM's embedding space, leaving paralinguistic information largely unexplored. Consequently, the utilization of paralinguistic information in current E2E ST systems is indirect and inadequate, leading to a suboptimal translation.

In this paper, we propose **P**ara**L**inguistic-**a**ware **S**peech **T**ranslation (**PLaST**), an end-to-end framework designed to enhance translation performance by directly leveraging paralinguistic information. PLaST comprises two parallel branches that independently capture linguistic and paralinguistic information. Specifically, the linguistic information is extracted and encoded by a speech encoder, while the paralinguistic features are obtained using a pre-trained style encoder tailored for paralinguistic tasks. To obtain purified linguistic content, we first align the speech and text modalities using hierarchical Optimal Transport (**OT**). In particular, the speech signals and corresponding transcript are fed into the LLM in pairs, and the alignment between the layer-wise decoder outputs of each modality is performed using the Wasserstein Loss (Frogner et al. 2015). This encourages consistent outputs regardless of input modality, effectively bridging the representational gap between speech and text. Importantly, the LLM is kept frozen during this process to ensure that the speech encoder is optimized independently. Next, we retrieve and refine the paralinguistic features via an Attention-based Retrieval (**AR**) module. This module employs a cross-attention (Vaswani et al. 2017) mechanism, where the aligned linguistic representations act as queries, and the paralinguistic features serve as keys and values, enabling effective paralinguistic integration. By seamlessly incorporating both linguistic and paralinguistic information into the LLM decoder, PLaST facilitates more comprehensive speech understanding and significantly improves translation quality. Experimental results indicate that paralinguistic information contributes substantially to the translation performance. PLaST achieves an average of **5.0** directional and **4.5** global contrastive likelihood scores improvements over the strong baseline on the paralinguistic-sensitive benchmark ContraProST. Furthermore, PLaST generalizes well to typical ST scenarios, where it outperforms the baseline with an average of 0.43 and 0.98 BLEU scores under 2B and 8B settings on the standard ST benchmark CoVoST-2.

Our contributions are summarized as follows:

- We propose **P**ara**L**inguistic-**a**ware **S**peech **T**ranslation (**PLaST**), a dual-branch framework in which paralinguistic cues beyond linguistic content are integrated into the LLM decoder to enhance semantic understanding and translation generation.
- We propose to apply a hierarchical Optimal Transport (**OT**) to enforce cross-modal alignment, thereby producing linguistically faithful representations. Moreover, the paralinguistic features derived by a style decoder are further retrieved and refined through the proposed Attention-based Retrieval (**AR**) module.
- Extensive experiments demonstrate that PLaST effectively leverages paralinguistic information, yielding substantial improvements over prior methods. It achieves state-of-the-art performance on both the paralinguistic awareness benchmark contraProST (Tsiamas et al. 2024b) and the general speech translation benchmark CoVoST-2 (Wang, Wu, and Pino 2020).

2 Methodology

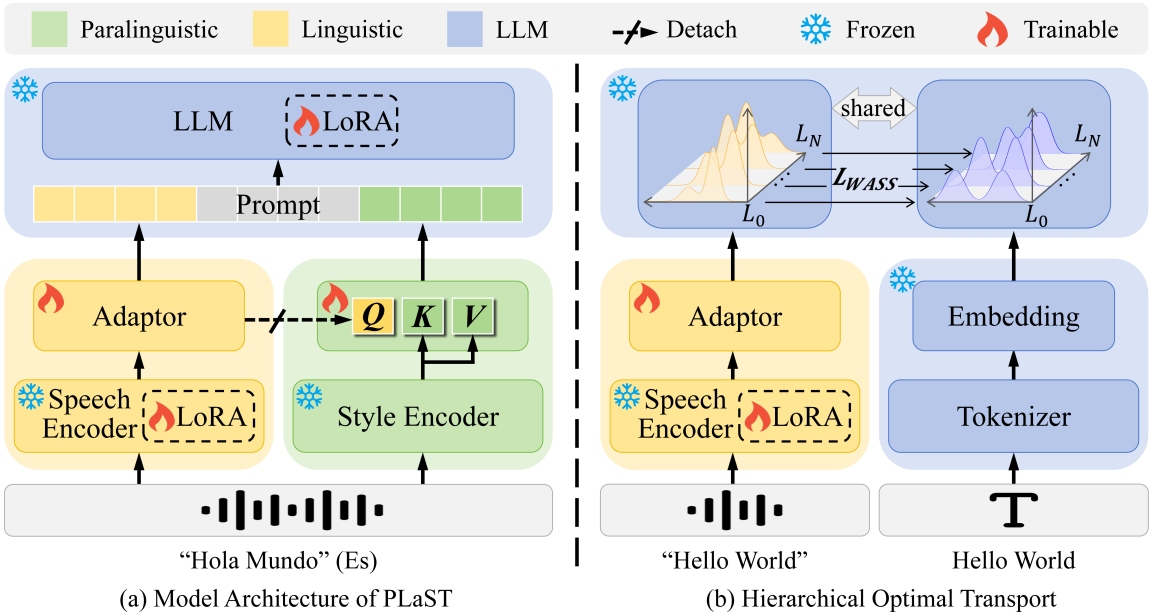
2.1 Model Architecture

Speech translation corpora typically comprise triplets $\{(s, x, y)\} \in \mathcal{D}_{\text{ST}}$, where s denotes the source speech, x is the source language transcript, and y is the target-language translation. E2E ST aims to directly generate y from s, eliminating the need for x as an intermediate step.

As depicted in Figure 2 (a), the proposed E2E ST framework PLaST consists of three components: (1) a *linguistic branch* comprising a speech encoder and an adaptor, (2) a *paralinguistic branch* including a style encoder and an Attention-based Retrieval module, and (3) an LLM decoder positioned on top of the whole framework for text generation. The two branches extract parallel linguistic and paralinguistic representations from the input speech. Subsequently, the retrieval module retrieves and refines the extracted paralinguistic representations, seamlessly integrating them into the LLM decoder to enhance translation performance.

2.2 The Linguistic Branch

Plentiful works have been proposed to extract the linguistic content in the source speech signals. Among them, Whisper (Radford et al. 2023) achieves state-of-the-art performance on diverse speech-related tasks due to its multi-task, large-scale, weakly supervised pre-training. Its effectiveness



(a) Model Architecture of PLaST (b) Hierarchical Optimal Transport

Figure 2: Overall of the proposed PLaST: (a) model architecture of PLaST, and (b) the cross-modal alignment with hierarchical Optimal Transport. The linguistic content and paralinguistic cues are separately extracted from distinct branches and fused with the proposed Attention-based Retrieval module, enabling an enhanced semantic understanding and translation generation.

has been validated in various E2E ST systems (Chen et al. 2024b; Huang et al. 2023).

We choose Whisper as the speech encoder to extract linguistic representation as (1) it exhibits strong multilingual generalization across languages and accents, (2) it is robust to real-world conditions such as background noise, speaker variability, and speech disfluencies, and (3) it provides a unified encoder architecture that is easy to integrate and fine-tune within E2E frameworks. Subsequently, a lightweight adaptor is utilized to project the speech embeddings into the dimension of the language model's input space:

$$H_l = \mathcal{F}^l_{\text{ada}}\big(\text{ENC}_{\text{speech}}(s)\big), \tag{1}$$

where $\text{ENC}_{\text{speech}}(\cdot)$ is the speech encoder, in particular, initialized with Whisper (Radford et al. 2023). The $\mathcal{F}_{\text{ada}}(\cdot)$ is a three-layer multilayer perceptron (MLP) that maps the encoder's output into the LLM decoder's embedding space.

Due to the inferior data quality during pre-training of Whisper and the informative nature of speech signals, the extracted linguistic representation might comprise undesired counterparts for translation. Thus, we propose to further align speech and text representations using Optimal Transport for a purified linguistic information extraction.

Optimal Transport. As depicted in Figure 2 (b), to generate linguistically faithful representations, we align the linguistic and text representations within the LLM's semantic space by minimizing the Wasserstein loss (Frogner et al. 2015) between the layer-wise decoder outputs of each paired input, which is grounded in OT (Peyré and Cuturi 2019). OT formulates transporting one distribution to another at minimum cost under a given cost matrix.

Let μ and ν be the empirical distributions of speech and text representations with length T_s and T_x, we have:

$$\begin{aligned} \mu &= \{(m_i, w_i)\}_{i=1}^{T_s}, \quad \text{s.t.} \;\; w_i = 1/T_s \\ \nu &= \{(m_j, w_j)\}_{j=1}^{T_x}, \quad \text{s.t.} \;\; w_j = 1/T_x \end{aligned} \tag{2}$$

where the uniform weights w_i and w_j are assigned to the elements m_i and m_j, respectively. Then, the squared Euclidean cost matrix $\mathbf{C} \in \mathbb{R}^{T_s \times T_x}$ is computed as $\mathbf{C}_{ij} = \|m_i - m_j\|_2^2$, measuring the transportation cost from m_i to m_j. Finally, the Wasserstein distance is defined as the minimal total transport cost:

$$\begin{aligned} &\mathcal{D}(\mu, \nu) = \min_{\mathbf{Z}} \sum_{i,j} \mathbf{Z}_{ij} \mathbf{C}_{ij}, \\ &\text{s.t.} \quad \sum_{i=1}^{T_s} \mathbf{Z}_{i:} = 1/T_s, \quad \forall i \in \{1, \dots, T_s\} \\ &\qquad \sum_{j=1}^{T_x} \mathbf{Z}_{:j} = 1/T_x, \quad \forall j \in \{1, \dots, T_x\} \end{aligned} \tag{3}$$

where $\mathbf{Z} \in \mathbb{R}^{T_s \times T_x}$ is the transport matrix that specifies how mass is transported between the two distributions.

In practice, we adopt the entropy-regularized upper-bound approximation of the Wasserstein distance to make it differentiable and efficiently computable by the Sinkhorn algorithm (Sinkhorn and Knopp 1967) following previous works (Tsiamas et al. 2024a; Le et al. 2023). Hence, the loss

function is defined as:

$$\mathcal{L}_{\text{WASS}} = \min_{\mathbf{Z}} \left(\sum_{i,j} \mathbf{Z}_{ij} \mathbf{C}_{ij} - \lambda \mathrm{H}(\mathbf{Z}) \right), \tag{4}$$

where $\mathrm{H}(\cdot)$ represents the von Neumann entropy, and $\lambda > 0$ controls the strength of the regularization.

To improve the robustness of alignment, we compute the Wasserstein loss $\mathcal{L}_{\text{WASS}}$ hierarchically across LLM layers:

$$\begin{aligned} \mu &= \text{LLM}_{\text{hidden}}\big(H_l\big), \\ \nu &= \text{LLM}_{\text{hidden}}\big(\text{EMB}(x')\big), \\ \mathcal{L} &= \sum_{k \in L} \alpha^k \, \mathcal{L}_{\text{WASS}}(\mu^k, \nu^k), \end{aligned} \tag{5}$$

where H_l is the speech representation derived from Equation (1), x' denotes the tokenized embedding of transcript x. L is the set of selected hidden layers in LLM, and α^k is the loss weight assigned to the k-th layer.

Note that the LLM remains frozen throughout this process to ensure the speech encoder is optimized in isolation, hence establishing a hierarchical alignment from speech to text.

2.3 The Paralinguistic Branch

In addition to linguistic content within the literal source transcript, the paralinguistic branch is meticulously designed to capture paralinguistic cues, thereby assisting the LLM decoder in end-to-end paralinguistic-aware translation. Therefore, a style encoder is employed to extract the paralinguistic features, including emphasis, emotion, prosody, et al.

Specifically, the emotion2vec (Ma et al. 2024) is adopted for this purpose due to its training-free character and fine-grained, frame-level feature extraction capacity. Symmetrically, we have:

$$Z_p = \text{ENC}_{\text{style}}(s). \tag{6}$$

However, since the representations extracted by the style encoder lack task-specific semantic alignment with the speech translation objective, the direct use of such features is suboptimal. We propose an Attention-based Retrieval mechanism to retrieve and refine the obtained paralinguistic representations for better adaptation to the ST task.

Attention-based Retrieval. The Attention-based Retrieval module is designed to retrieve and integrate paralinguistic information into the translation process. In this module, the linguistic representations extracted by the speech encoder serve as queries, while the paralinguistic features derived from the style encoder are used as keys and values:

$$H_p = \text{Attention}\big(\text{Detach}(H_l),\, Z_p,\, Z_p\big). \tag{7}$$

Here, H_l is detached in advance to prevent gradients from flowing back into the linguistic branch, thereby preserving the integrity of its semantically aligned representation space. This design enables the model to selectively retrieve paralinguistic cues most relevant to the ST task.

In contrast to the subsequent self-attention operations within the LLM decoder, this retrieval mechanism employs a decoupled query and key/value formulation, similar to the encoder-decoder attention found in standard sequence-to-sequence models (Vaswani et al. 2017). This architectural choice promotes a structured fusion of linguistic and paralinguistic information, facilitating precise and effective integration of emotional, prosodic, and emphatic cues into the translation process. By explicitly separating linguistic and paralinguistic from distinct branches and fusing them with our proposed AR module at the attention level, the model is better equipped to perceive, adapt to, and utilize auxiliary paralinguistic signals, thereby enhancing the semantic and pragmatic adequacy of the generated translation.

Prompt for ASR:

<H_l> Please **transcribe** the above **English** sentence into **English**, given the following auxiliary information. <H_p>

Prompt for ST:

<H_l> Please **translate** the above **Spanish** sentence into **English**, given the following auxiliary information. <H_p>

Figure 3: An example of the prompt template. In bold are instructive words that differ across tasks and translation directions. <H_l> is the linguistic features, and <H_p> denotes the paralinguistic representations extracted by the linguistic and paralinguistic branches, respectively.

2.4 The LLM Decoder

Given the obtained linguistic representation H_l and the paralinguistic representation H_p, the LLM decoder aims to generate high-fidelity translations. We adopt Llama2 for its strong performance in text generation (Zhang et al. 2024a; Touvron et al. 2023).

To leverage both types of information, we perform instruction tuning (Chen et al. 2024b; Zhang et al. 2023b) by integrating H_l and H_p into a unified instruction. As illustrated in Figure 3, the instruction contains a task-specific prompt, the linguistic representation H_l, and the paralinguistic representation H_p. Finally, with these three components concatenated as the LLM input, we have the training objective:

$$p(y|s) = \prod_{i=1} p(y_i | H_l, \text{Prompt}, H_p, y_{<i}). \tag{8}$$

3 Experiments

3.1 Experimental Settings

Datasets. ***ContraProST*** (Tsiamas et al. 2024b) is a paralinguistic benchmark which comprises double-contrastive examples, each consisting of a potentially semantically ambiguous English transcript paired with two distinct <speech, translation> pairs that differ in paralinguistic expression. ***CoVoST-2*** (Wang, Wu, and Pino 2020) is a large-scale multilingual speech translation corpus derived from Common Voice (Ardila et al. 2019). Following previous works (Chen et al. 2024b), we utilize 9 directions of data for training, involving *Fr/De/Es/It/Zh/Ja* → *En* and *En* → *De/Ja/Zh*,

Model	German		Spanish		Japanese	
	Directional	Global	Directional	Global	Directional	Global
XLS-R 2B (2022)	59.3	4.6	57.6	5.6	60.0	4.6
SeamlessM4T (large-v2) (2023)	61.2	13.5	64.9	13.4	59.4	12.4
ZeroSwot-Large (2024a)	60.6	9.7	57.5	9.2	58.8	7.9
SALMONN-13B (2024)	62.8	7.2	61.3	3.6	60.4	10.8
LLaST-2B† (2024b)	63.2	10.2	60.4	10.8	56.3	8.7
PLaST-2B (ours)	**65.5** (2.3↑)	**15.4** (5.2↑)	**66.3** (5.9↑)	**13.6** (3.8↑)	**63.1** (6.8↑)	**13.3** (4.6↑)

Table 1: Main results on the paralinguistic-sensitive benchmark ContraProST, including 3 directions from English → X. Directional and Global donate corresponding contrastive likelihood scores. †: our reproduction. The optimal results are highlighted in bold. Our performance improvements over the Baseline LLaST (Chen et al. 2024b) are also displayed for a clear comparison.

Model	Parms.	French	Japanese	German	Chinese	Spanish	Italian
Whisper-large-v2 (2023)	1.6B	36.4	26.1	36.3	18.0	40.1	30.9
SeamlessM4T (medium) (2023)	1.2B	38.4	15.2	34.7	18.0	38.7	36.5
SeamlessM4T (large-v2) (2023)	2.3B	42.1	23.8	39.9	22.2	42.9	40.0
SpeechLLaMA (2023)	7B	25.2	19.9	27.1	12.3	27.9	25.9
Qwen-audio (2023)	8B	38.5	-	33.9	15.7	39.7	36.0
LLaST-2B (2024b)	2B	41.2	24.2	36.8	19.2	43.2	39.3
LLaST-8B (2024b)	8B	44.1	24.4	40.8	23.3	45.3	42.1
PLaST-2B (ours)	2B	41.6 (0.4↑)	24.8 (0.6↑)	37.4 (0.6↑)	19.7 (0.5↑)	43.3 (0.1↑)	39.7 (0.4↑)
PLaST-8B (ours)	8B	**44.3** (0.2↑)	**27.8** (3.4↑)	**40.9** (0.1↑)	**24.8** (1.5↑)	**45.5** (0.2↑)	**42.6** (0.5↑)

Table 2: Main results on the standard speech translation benchmark CoVoST-2, including 6 directions from X → English.

selected from a total pool of 21 languages translated into English and 15 languages translated from English.

Evaluation Metrics. The *contrastive evaluation* (Sennrich 2017), which measures how well an ST model handles paralinguistic information, is adopted as the primary metric on ContraProST. Specifically, we use contrastive likelihood and contrastive quality (Tsiamas et al. 2024b) to assess the model's paralinguistic awareness. While on CoVoST-2, a case-sensitive sacreBLEU (Post 2018) is utilized to evaluate translation accuracy.

Model Configurations. In the linguistic branch, Whisper-large-v2 (Radford et al. 2023) is employed as the speech encoder, and the adaptor is a three-layer MLP that projects 1280-dimensional representations into the LLM's 2048-dimensional embedding space. In the paralinguistic branch, the frozen emotion2vec+base (Ma et al. 2024) is set as the style encoder to extract paralinguistic cues. The AR module is implemented via separate linear layers for queries, keys, values, and outputs, followed by a three-layer MLP. To ensure a fair comparison with our baseline (Chen et al. 2024b), TinyLlama-1.1B-Chat (Zhang et al. 2024a) and Llama-2-7b-chat (Touvron et al. 2023) were used as the LLM decoder, resulting in a total of approximately 2B and 8B parameters, respectively. For efficient training, we apply LoRA (Hu et al. 2022) adaptors to the speech encoder and the LLM, configuring the rank to 128 for the former and 512 for the latter.

Training Details. The alignment between speech and text representations with the proposed hierarchical Optimal Transport is first conducted for linguistically faithful information extraction. Particularly, the speech encoder and adaptor in the linguistic branch are optimized by minimizing $\mathcal{L}_{\text{WASS}}$ across layer-wise decoder outputs in LLM with $k \in \{5, 7, 9, 11, 13, 15, 17, 19, 21, 22\}$, and the corresponding weights $\alpha^k \in \{0.4, 0.4, 0.4, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. Subsequently, an E2E ST is performed to train the entire model, where the trainable parameters include the adaptor, the AR module, and LoRA adaptors for both the speech encoder and the LLM decoder. The alignment and E2E training are separately performed for one epoch using the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate warms up linearly over the first 3% of training steps to a peak of 2×10^{-4}, followed by cosine annealing.

3.2 Main Results

ContraProST. As illustrated in Table 1, we present the results of our proposed PLaST on the ContraProST benchmark in all three directions translated from English: German, Spanish, and Japanese. Apart from the baseline LLaST (Chen et al. 2024b), we also report other representative models as follows: XLS-R (Babu et al. 2022), an E2E model built on WAV2VEC2.0 and mBART; SeamlessM4T (Communication et al. 2023), a multilingual and multimodal encoder-decoder model trained with multi-task learning on ASR, MT, and ST; ZeroSwot-Large (Tsiamas et al. 2024a), a zero-shot E2E ST model that combines WAV2VEC2.0 with an NLLB decoder; and SALMONN (Tang et al. 2024), an audio LLM that integrates

Model	Lin.	Para.	OT	AR	Detach	Likelihood	BLEU
A	✓	✗	✓	-	-	54.3	41.38
B	✓	✓	✗	✓	✓	61.8	41.43
C	✓	✓	✓	✗	✓	56.2	41.49
D	✓	✓	✓	✓	✗	57.0	41.52
PLaST	✓	✓	✓	✓	✓	**62.9**	**41.80**

Table 3: Study on the contribution of each component. Results are reported on CoVoST-2 for *Es→En* and on ContraProST for *En→Es* contrastive likelihood.

whisper and BEATs to the Vicuna decoder.

Due to the nature of the E2E ST method, where it has direct access to the speech signals, all models possess at least a partial internal representation of paralinguistic cues, which allows them to outperform the random baseline of 50% in the directional contrastive likelihood. On the other hand, correctly solving both sub-cases of each example (i.e., the global contrastive likelihood) presents a highly challenging objective, as performance scores rarely exceed 10%, except in the case of SeamlessM4T, which utilizes an unsupervised pretraining strategy, potentially leading to a broader and diverse acoustic speech representation.

In contrast, PLaST consistently outperforms all these methods by a substantial margin in both directional and global contrastive likelihood due to its direct exploitation of paralinguistic information. Particularly, PLaST achieves average scores of 65.0 and 14.1 on directional and global likelihood, an improvement over LLaST with gains ranging from 2.3 to 5.9 points in directional likelihood and 2.8 to 5.2 points in global likelihood. These results highlight PLaST's exceptional ability to leverage both linguistic and paralinguistic information, enabling fine-grained translation decisions that outperform previous state-of-the-art models.

CoVoST-2. To further generalize the proposed PLaST to a more typical ST scenario, we evaluate our model on the standard ST benchmark CoVoST-2. In line with the baseline PLaST (Chen et al. 2024b), training data from 9 translation directions are merged to build a unified model and BLEU scores are reported independently on 6 X to English translation directions: *Fr/De/Es/It/Zh/Ja → En*. The results are shown in Table 2. PLaST-2B achieves an average improvement of 0.43 BLEU scores over LLaST, and it demonstrates competitive performance with SeamlessM4T-Large-v2, while attaining the best result on translating from Spanish to English (*Es → En*) among methods of comparable parameters. PLaST-8B outperforms all previous methods and achieves new state-of-the-art performance at a similar scale, with a notable 3.4 BLEU score improvement on the *Ja → En*, where relatively low Japanese parallel training data is involved during training. These results demonstrate that PLaST generalizes well to general translation scenarios and underscore the substantial benefits of paralinguistic augmentation, especially for low-resource languages.

Model	Top-1 Accuracy		
	22	21	20
LLaST-2B	91.3	66.7	21.2
PLaST-2B	**97.5**	**97.3**	**96.9**

Table 4: Cross-modal retrieval accuracy on CoVoST-2 *Es→En* test set. The Top-1 accuracy is evaluated using the Wasserstein distance between speech and text representations from the LLM decoder layers.

3.3 Ablation Studies

Study on the Contribution of Each Component. An ablation study is conducted by removing the component to evaluate its contribution to our model, as shown in Table 3.

Firstly, the removal of the entire paralinguistic branch leads to the most significant decline of 8.6 and 0.42 in the likelihood and the BLEU scores (in line A), which demonstrates the effectiveness of paralinguistic cues in enhancing the translation process. The OT and AR modules contribute to the paralinguistic-aware capability of PLaST to a different extent, which improves the likelihood score by 1.1 and 6.7 and the BLEU score by 0.37 and 0.31, respectively (in lines B and C). Finally, we remove the detach operations in AR, where the gradients could flow back to the linguistic branch without a hitch, resulting in a decline of 5.9 in the likelihood and 0.28 in the BLEU scores (in line D). This suggests that information from the paralinguistic branch may interfere with the linguistic representation space, and that explicitly preserving the separation between linguistic and paralinguistic information through distinct branches contributes to a better speech understanding in the model.

Study on the Effectiveness of OT. We conduct a retrieval experiment to assess whether the proposed hierarchical OT effectively aligns source speech with corresponding textual content and promotes the generation of linguistically faithful representations. Specifically, 2,000 <speech, transcript> pairs are randomly sampled from the of the *Es → En* test set on CoVoST-2. For each speech input, we identify its nearest transcript neighbor using the minimal Wasserstein distance and calculate accuracy on these samples. Retrieval is conducted on the final three layers of the LLM decoder.

The results in Table 4 indicate that the proposed PLaST takes a dominant lead over the baseline LLaST. Though both PLaST and LLaST achieve high accuracy on the final layer outputs, i.e., exceeding 90% accuracy on the 22nd layer, our PLaST still outperforms LLaST by an absolute margin of 6.2%. When moving to lower layers, a clear divergence emerges where LLaST's accuracy drops sharply to 66.7% and 21.2% at the 21st and 20th layer, a huge decline of 24.6% and 70.1%. However, PLaST maintains a high accuracy of 97.3% and 96.9%, with a marginal decrease of 0.2% and 0.6% in accuracy, respectively. These findings highlight the effectiveness of Optimal Transport in aligning source speech with corresponding textual content, enabling the generation of linguistically faithful speech representations.

Model	Likelihood	BLEU
Self-attention on paralinguistics	60.7	41.68
Self-attention on linguistics	54.6	41.55
w/ 1-layer AR	62.9	41.80
w/ 2-layer AR	62.4	41.75
w/ 3-layer AR	62.5	41.79

Table 5: Study on the AR variations in PLaST. The *Es* $\rightarrow$ *En* subset in CoVoST-2 is used for training, and the global contrastive likelihood is evaluated on ContraProST.

Study on the Effectiveness of AR. To explore the role of the proposed AR in PLaST, we select several variations of AR to analyze the Effectiveness of AR, as shown in Table 5.

We first replace the Retrieval module implemented based on cross-attention with self-attention. In this situation, the Retrieval module degenerates to a vanilla feature extractor. Taking the extracted paralinguistic representation as input, the non-retrieved and full paralinguistic information is reserved during the translation process, which degrades the likelihood and BLEU scores by 2.2 and 0.12. In another variation, we replace the input with the linguistic content, excluding the paralinguistic cues, to further investigate the effectiveness of paralinguistic information. This setup can be viewed as enhancing translation by providing linguistic information to the decoder twice. Interestingly, a slight improvement in BLEU score is observed compared to the non-enhanced variation(see Table 3, line A), while the likelihood score remains quite the same. This confirms that the speech encoder, despite having direct access to the speech signals, primarily captures textual content. The performance of the other three variations with stacked cross-attention layers shows little difference, indicating that the improvement stems from the retrieval process rather than the increase in model capacity. Thus, the one-layer AR is adopted in PLaST.

4 Related Work

4.1 End-to-End Speech Translation

E2E ST directly translates source speech signals into target text, eliminating the need for an intermediate transcript. Existing methods treat ST as a cross-modal task, aiming to bridge the modality gap between source speech and text. Typically, Tang et al. (2021) and Ye, Wang, and Li (2021) proposed bridging the representation discrepancy implicitly by projecting speech and text into a shared semantic space within a multi-task learning framework. Meanwhile, various approaches have been proposed to explicitly align speech and text representations using auxiliary optimization techniques, such as Optimal Transport (Tsiamas et al. 2024a), contrastive learning (Ouyang, Ye, and Li 2023; Ye, Wang, and Li 2022), and adversarial training (Zhang et al. 2024b). The recent advances of LLMs in semantic understanding and language modeling further pushed the E2E ST ahead (Gaido et al. 2024; Chen et al. 2024b; Tang et al. 2024; Zhang et al. 2023a; Wu et al. 2025), where an adaptor is typically employed to fit the LLM's embedding space. Although E2E ST models appear well-suited for paralinguistic-aware translation due to their direct access to speech signals, overly aligned representations may hinder the models' capability to leverage paralinguistic information, as it pushes the models to treat speech signals closely aligned with their linguistic content.

Distinguished from previous E2E ST methods that focus solely on linguistic information, the proposed PLaST leverages paralinguistic cues beyond the linguistic part via a dual-branch encoder, capitalizing on the advancements of LLMs. This enables a more comprehensive understanding of speech signals and an enhanced translation quality.

4.2 Paralinguistics

Paralinguistic cues have attracted sustained and widespread interest across various speech-related tasks. Skerry-Ryan et al. (2018) proposed to learn a latent representation of prosody to transfer in text-to-speech synthesis (TTS), while Pamisetty, Sri, and Murty (2021) incorporated desired prosodic features by controlling the fundamental frequency and the phone duration in TTS. In the speech-to-speech translation (S2S) domain, Communication et al. (2023) and Do, Sakti, and Nakamura (2017) have discussed the expressiveness of the translated speech. Furthermore, paralinguistic cues have been integrated into the pre-training of LLMs, leading to the development of multi-modal LLMs for enhanced speech-based question answering and dialogue generation (Chen et al. 2025; Lin, Chiang, and Lee 2024).

In the ST domain, Zhou et al.(2024) introduced contrastive likelihood to study the paralinguistic-aware capability of E2E and cascaded ST models in Korean wh-phrases, and empirically demonstrated that E2E models better utilize paralinguistic information. Meanwhile, Tsiamas et al.(2024b) contributed ContraProST, a double-contrastive dataset designed to facilitate a broader study of paralinguistic phenomena in speech translation, offering us a reliable benchmark for evaluating paralinguistic awareness. Closely related, Chen et al.(2024a) proposed MELD-ST, which used emotion labels to boost translation performance and constructed an emotion-aware speech translation dataset. However, their approach relies on simple, coarse-grained emotion labels, which are often limited in availability. In contrast, we propose the utilization of refined, frame-level paralinguistic representations, which contribute to a fine-grained paralinguistic-aware speech translation system.

5 Conclusion

E2E ST methods currently face limitations in effectively utilizing paralinguistic information. Our proposed PLaST mitigates this gap by directly leveraging paralinguistic cues through a dual-branch framework. By independently extracting linguistic and paralinguistic representations, applying hierarchical Optimal Transport for purified linguistic representation generation, and using an Attention-based Retrieval module for paralinguistic information refinement, PLaST showcases its strength on the paralinguistic-sensitive benchmark ContraProST, and it generalizes well to the standard speech translation benchmark CoVoST-2.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grants 62476232, 62036004) and the University-Industry Cooperation Program of Fujian Province (Grant 2023H6001).

References

Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F. M.; and Weber, G. 2019. Common Voice: A Massively-Multilingual Speech Corpus. *ArXiv*, abs/1912.06670.

Babu, A.; Wang, C.; Tjandra, A.; Lakhotia, K.; Xu, Q.; Goyal, N.; Singh, K.; von Platen, P.; Saraf, Y.; Pino, J.; Baevski, A.; Conneau, A.; and Auli, M. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Interspeech 2022*, 2278–2282.

Banse, R.; and Scherer, K. R. 1996. Acoustic Profiles in Vocal Emotion Expression. *Journal of Personality and Social Psychology*, 70(3): 614–636.

Bolinger, D. L. 1961. Contrastive Accent and Contrastive Stress. *Language*, 37(1): 83.

Castro, S.; Hazarika, D.; Pérez-Rosas, V.; Zimmermann, R.; Mihalcea, R.; and Poria, S. 2019. Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper). In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4619–4629. Florence, Italy: Association for Computational Linguistics.

Chen, H.; Li, Z.; Song, Y.; Deng, W.; Yao, Y.; Zhang, Y.; Lv, H.; Zhu, X.; Kang, J.; Lian, J.; Li, J.; Wang, C.; Song, S.; Li, Y.; He, Z.; and Li, X. 2025. GOAT-SLM: A Spoken Language Model with Paralinguistic and Speaker Characteristic Awareness. arXiv:2507.18119.

Chen, S.; Yahata, S.; Shimizu, S.; Yang, Z.; Li, Y.; Chu, C.; and Kurohashi, S. 2024a. MELD-ST: An Emotion-aware Speech Translation Dataset. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 10118–10126. Bangkok, Thailand: Association for Computational Linguistics.

Chen, X.; Zhang, S.; Bai, Q.; Chen, K.; and Nakamura, S. 2024b. LLaST: Improved End-to-end Speech Translation System Leveraged by Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 6976–6987. Bangkok, Thailand: Association for Computational Linguistics.

Chu, Y.; Xu, J.; Zhou, X.; Yang, Q.; Zhang, S.; Yan, Z.; Zhou, C.; and Zhou, J. 2023. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. *ArXiv*, abs/2311.07919.

Communication, S.; Barrault, L.; Chung, Y.-A.; Meglioli, M. C.; Dale, D.; Dong, N.; Duppenthaler, M.; Duquenne, P.-A.; Ellis, B.; ElSahar, H.; Haaheim, J.; Hoffman, J.; Hwang, M.-J.; Inaguma, H.; Klaiber, C.; Kulikov, I.; Li, P.; Licht, D.; Maillard, J.; Mavlyutov, R.; Rakotoarison, A.; Sadagopan, K. R.; Ramakrishnan, A.; Tran, T.; Wenzek, G.; Yang, Y.; Ye, E.; Evtimov, I.; Fernandez, P.; Gao, C.; Hansanti, P.; Kalbassi, E.; Kallet, A.; Kozhevnikov, A.; Gonzalez, G. M.; Roman, R. S.; Touret, C.; Wong, C.; Wood, C.; Yu, B.; Andrews, P.; Balioglu, C.; Chen, P.-J.; Costa-jussà, M. R.; Elbayad, M.; Gong, H.; Guzm'an, F.; Heffernan, K.; Jain, S.; Kao, J. T.; Lee, A.; Ma, X.; Mourachko, A.; Peloquin, B.; Pino, J.; Popuri, S.; Ropers, C.; Saleem, S.; Schwenk, H.; Sun, A.; Tomasello, P.; Wang, C.; Wang, J.; Wang, S.; and Williamson, M. 2023. Seamless: Multilingual Expressive and Streaming Speech Translation. *ArXiv*, abs/2312.05187.

Do, Q. T.; Sakti, S.; and Nakamura, S. 2017. Toward Expressive Speech Translation: A Unified Sequence-to-Sequence LSTMs Approach for Translating Words and Emphasis. In *Interspeech 2017*, 2640–2644.

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)*, 15864–15881. Toronto, Canada: Association for Computational Linguistics.

Fang, Q.; Ye, R.; Li, L.; Feng, Y.; and Wang, M. 2022. STEMM: Self-learning with Speech-text Manifold Mixup 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)*, 7050–7062. Dublin, Ireland: Association for Computational Linguistics.

Frogner, C.; Zhang, C.; Mobahi, H.; Araya-Polo, M.; and Poggio, T. 2015. Learning with a Wasserstein loss. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, 2053–2061. Cambridge, MA, USA: MIT Press.

Gaido, M.; Papi, S.; Negri, M.; and Bentivogli, L. 2024. Speech Translation with Speech Foundation Models and Large Language Models: What is There and What is Missing? 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)*, 14760–14778. Bangkok, Thailand: Association for Computational Linguistics.

Han, C.; Wang, M.; Ji, H.; and Li, L. 2021. Learning Shared Semantic Space for Speech-to-Text Translation. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2214–2225. Online: Association for Computational Linguistics.

Hu, E. J.; yelong shen; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, Z.; Ye, R.; Ko, T.; Dong, Q.; Cheng, S.; Wang, M.; and Li, H. 2023. Speech Translation with Large Language Models: An Industrial Practice. *ArXiv*, abs/2312.13585.

Le, P.-H.; Gong, H.; Wang, C.; Pino, J.; Lecouteux, B.; and Schwab, D. 2023. Pre-training for speech translation: CTC meets optimal transport. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Lin, G.-T.; Chiang, C.-H.; and Lee, H.-y. 2024. Advancing Large Language Models to Capture Varied Speaking Styles and Respond Properly in Spoken Conversations. 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)*, 6626–6642. Bangkok, Thailand: Association for Computational Linguistics.

Ma, Z.; Zheng, Z.; Ye, J.; Li, J.; Gao, Z.; Zhang, S.; and Chen, X. 2024. emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 15747–15760. Bangkok, Thailand: Association for Computational Linguistics.

Ouyang, S.; Ye, R.; and Li, L. 2023. WACO: Word-Aligned Contrastive Learning 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), 3891–3907. Toronto, Canada: Association for Computational Linguistics.

Pamisetty, G.; Sri, K.; and Murty, R. 2021. Prosody-TTS: An End-to-End Speech Synthesis System with Prosody Control. *Circuits, Systems, and Signal Processing*, 42: 361–384.

Peyré, G.; and Cuturi, M. 2019. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607.

Post, M. 2018. A Call for Clarity in Reporting BLEU Scores. In *Conference on Machine Translation*.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; Mcleavey, C.; and Sutskever, I. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 28492–28518. PMLR.

Sennrich, R. 2017. How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In Lapata, M.; Blunsom, P.; and Koller, A., eds., *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 376–382. Valencia, Spain: Association for Computational Linguistics.

Shriberg, E.; Stolcke, A.; Jurafsky, D.; Coccaro, N.; Meteer, M.; Bates, R.; Taylor, P.; Ries, K.; Martin, R.; and Van Ess-Dykema, C. 1998. Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? *Language and Speech*, 41(3-4): 443–492.

Sinkhorn, R.; and Knopp, P. 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21: 343–348.

Skerry-Ryan, R.; Battenberg, E.; Xiao, Y.; Wang, Y.; Stanton, D.; Shor, J.; Weiss, R.; Clark, R.; and Saurous, R. A. 2018. Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 4693–4702. PMLR.

Sperber, M.; and Paulik, M. 2020. Speech Translation and the End-to-End Promise: Taking Stock of Where We Are. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7409–7421. Online: Association for Computational Linguistics.

Tang, C.; Yu, W.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; MA, Z.; and Zhang, C. 2024. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Tang, Y.; Pino, J.; Wang, C.; Ma, X.; and Genzel, D. 2021. A General Multi-Task Learning Framework to Leverage Text Data for Speech to Text Tasks. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6209–6213.

Touvron, H.; Martin, L.; Stone, K. R.; Albert, P.; Almahairi, A.; Babaei, Y.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Tsiamas, I.; Gállego, G. I.; Fonollosa, J. A. R.; and Costa-jussà, M. R. 2024a. 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*, 14245–14267. Bangkok, Thailand: Association for Computational Linguistics.

Tsiamas, I.; Sperber, M.; Finch, A.; and Garg, S. 2024b. Speech Is More than Words: Do Speech-to-Text Translation Systems Leverage Prosody? In Haddow, B.; Kocmi, T.; Koehn, P.; and Monz, C., eds., *Proceedings of the Ninth Conference on Machine Translation*, 1235–1257. Miami, Florida, USA: Association for Computational Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30, 6000–6010. Curran Associates, Inc.

Wang, C.; Wu, A.; and Pino, J. M. 2020. CoVoST 2 and Massively Multilingual Speech-to-Text Translation. *arXiv: Computation and Language*.

Wu, J.; Gaur, Y.; Chen, Z.; Zhou, L.; Zhu, Y.; Wang, T.; Li, J.; Liu, S.; Ren, B.; Liu, L.; and Wu, Y. 2023. On Decoder-Only Architecture For Speech-to-Text and Large Language Model Integration. *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1–8.

Wu, S.; Tang, J.; Yang, C.; Zhang, P.; Yang, B.; Li, J.; Yao, J.; Zhang, M.; and Su, J. 2025. Locate-and-Focus: Enhancing Terminology Translation in Speech Language Models. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11345–11360. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Ye, R.; Wang, M.; and Li, L. 2021. End-to-End Speech Translation via Cross-Modal Progressive Training. In *Interspeech 2021*, 2267–2271.

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*, 5099–5113. Seattle, United States: Association for Computational Linguistics.

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. *ArXiv*, abs/2310.02050.

Zhang, P.; Zeng, G.; Wang, T.; and Lu, W. 2024a. TinyLlama: An Open-Source Small Language Model. *ArXiv*, abs/2401.02385.

Zhang, S.; Dong, L.; Li, X.; Zhang, S.; Sun, X.; Wang, S.; Li, J.; Hu, R.; Zhang, T.; Wu, F.; and Wang, G. 2023b. Instruction Tuning for Large Language Models: A Survey. *ArXiv*, abs/2308.10792.

Zhang, Y.; Kou, K.; Li, B.; Xu, C.; Zhang, C.; Xiao, T.; and Zhu, J. 2024b. Soft Alignment of Modality Space for End-to-End Speech Translation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 11041–11045.

Zhou, G.; Lam, T. K.; Birch, A.; and Haddow, B. 2024. Prosody in Cascade and Direct Speech-to-Text Translation: a case study on Korean Wh-Phrases. In Graham, Y.; and Purver, M., eds., *Findings of the Association for Computational Linguistics: EACL 2024*, 674–683. St. Julian's, Malta: Association for Computational Linguistics.

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)*, 7873–7887. Toronto, Canada: Association for Computational Linguistics.