

PART: PROGRESSIVE ALIGNMENT REPRESENTATION TRAINING FOR MULTILINGUAL SPEECH-TO-TEXT WITH LLMs

*Pei Zhang^{*13}, Andong Chen^{*1}, Xi Chen^{*12}, Baosong Yang¹, Derek F. Wong³, Fei Huang¹*

¹ Tongyi Lab, Alibaba Group, ²The Chinese University of Hong Kong,
³ NLP²CT Lab, University of Macau

ABSTRACT

Large language models (LLMs) have expanded from text to speech, giving rise to Speech Large Models (SLMs) that support recognition, translation, and synthesis. A key challenge is aligning speech and text representations, which becomes harder in multilingual settings. Existing methods often freeze LLM parameters and train encoders on multilingual data, but this forces cross-language convergence and limits performance. We introduce Progressive Alignment Representation Training (PART), a multi-stage and multi-task framework that separates within-language from cross-language alignment. During cross-language training, LLM parameters are dynamically activated, and text-based tasks are later introduced to enhance multilingual understanding. Experiments on CommonVoice 15, Fleurs, Wenetspeech, and CoVoST2 show that PART surpasses conventional approaches, with analysis confirming its ability to balance language-specific distinctions and cross-language generalization. These results demonstrate PART’s effectiveness and generality for multilingual speech modality alignment.

Index Terms— Multilingual Speech Processing, Speech-Text Alignment, Large Language Models

1. INTRODUCTION

In the research of large language models (LLMs), the scope has expanded from text to other modalities. Among them, Speech Large Models (SLMs) that use speech input and output have been widely studied and applied, showing impressive performance in tasks such as speech recognition [1], translation [2, 3], and speech synthesis [4].

The mainstream architecture of current SLMs usually consists of a pre-trained speech encoder connected to an LLM through an adapter [5, 6]. Within this framework, a central challenge is how to effectively align speech representations with the textual representations of the LLM [7, 8]. This need becomes more critical and difficult in fine-grained multilingual scenarios. In practice, the conventional approach often treats multilingual tasks as monolingual ones, mixing non-crosslingual tasks like automatic speech recognition (ASR) with crosslingual tasks like speech to text translation (S2TT) during training. A common approach is to keep the LLM parameters frozen and train the speech encoder on multilingual speech data so that it aligns with the LLM input layer [9, 10]. However, when applied to multilingual tasks, this strategy may force audio representations of different languages to converge and place an excessive burden on the speech encoder [11, 12]. As a result, such methods remain at the modality-level alignment, lacking

These authors contributed equally to this work and should be considered co-first authors.

deep exploration of multilingual speech-text alignment and facing performance bottlenecks in multilingual speech tasks [13, 14].

We propose a Progressive Alignment Representation Training approach (**PART**) for Multilingual SLMs. In this approach, the training is divided into stages to separate within-language alignment from cross-language alignment, preventing excessive convergence of audio representations across languages. For cross-language tasks, LLM parameters are dynamically activated, allowing the speech encoder to focus on semantic mapping within each language while the LLM leverages its multilingual modeling strength. In the final stage, text-based tasks are introduced to fine-tune the LLM, further improving its ability in multilingual instruction understanding and generation. Overall, our method enables a collaborative division of labor between the speech encoder and the LLM, preserving language-specific distinctions while enhancing cross-language generalization.

Compared with traditional training methods, our multi-stage and multi-task alignment approach achieves better performance on CommonVoice 15, Fleurs, and Wenetspeech (ASR tasks), as well as CoVoST2 (S2TT task). In addition, analytical experiments show that activating the LLM in cross-language tasks effectively leverages its multilingual modeling strength, while introducing text-based tasks for fine-tuning in the final stage significantly improves multilingual ability. Overall, the results confirm the effectiveness and generality of our method for multilingual speech modality alignment.

Our contributions are summarized as follows:

- We propose a multi-stage, multi-task alignment framework for multilingual speech-LLMs, which separates in-language alignment from multilingual alignment to better preserve language-specific features.
- We design a task-dependent activation strategy that freezes the LLM in in-language ASR tasks while activating it in multilingual tasks (e.g., S2TT), allowing the audio encoder and LLM to play to their respective strengths.
- We further introduce a final text-based fine-tuning stage that enhances multilingual ability, leading to improved performance across ASR and S2TT tasks.

2. METHOD

In this section, we first formulate the multilingual speech-to-text translation task in 2.1, then present the model architecture of our proposed method in 2.2, followed by a detailed description of our progressively aligned training strategy in 2.3.

2.1. Problem Formulation

We formulate the multilingual speech-to-text task as follows: given a speech input x , the goal is to generate its corresponding textual

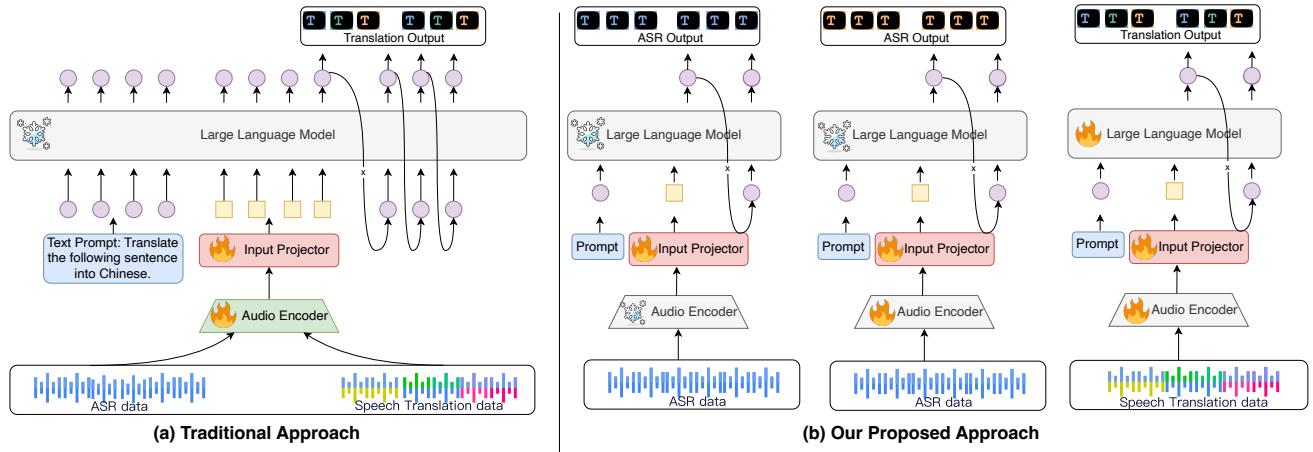


Fig. 1. Our proposed training recipe v.s. traditional training approach

output y . This work focuses on two distinct settings: multilingual ASR and S2TT. The language of x is referred to as the speech in source language. For multilingual ASR, which is a monolingual task, the output y is a transcription in the same language as x , trained using the monolingual dataset $\mathcal{D}_{\text{mono}}$. In contrast, speech translation is a cross-lingual task where y is produced in a target language different from the source language, leveraging the cross-lingual dataset $\mathcal{D}_{\text{cross}}$.

2.2. Model Architecture

As in [15, 16], we first extract log-mel spectrogram features \mathbf{M} from the input speech signal \mathbf{X} using a feature extractor. These features are then fed into a pre-trained multilingual speech encoder (parameterized by θ_{se}) to obtain linguistic representations \mathbf{X}_{ling} . Subsequently, a lightweight adaptor module, randomly initialized with parameters θ_{adaptor} and not pretrained, projects \mathbf{X}_{ling} into the embedding space of a large language model (LLM, parameterized by θ_{llm}), resulting in the aligned feature representation $\mathbf{X}_{\text{align}}$. This adapted representation is then concatenated with the embedding of the instruction token \mathbf{X}_I , and the combined input is passed to the multilingual LLM. Finally, the LLM generates the corresponding text. Method details are shown in Figure 1.

2.3. Progressive Training for Multilingual SLMs

In contrast to existing methods, this paper contends that multilingual tasks introduce significant challenges for alignment. To address this, a progressive alignment approach is adopted, consisting of three main stages. The first two stages are responsible for gradual cross-modality alignment, while the third stage builds upon this modality-aligned foundation to fine-tune on cross-lingual tasks, thereby better leveraging the multilingual capabilities of the LLM. All three stages share the same underlying optimization objective, which is formally defined by the loss function in Eq. 1

$$\begin{aligned} \mathcal{L}_{\mathcal{D}} &= -\mathbb{E}_{(\mathbf{X}, \mathbf{y}) \sim \mathcal{D}} \log P(\mathbf{y} | \mathbf{X}; \theta_{\text{se}}, \theta_{\text{adaptor}}, \theta_{\text{llm}}) \\ P(\mathbf{y} | \mathbf{X}) &= \prod_{t=1}^T P(y_t | \mathbf{y}_{<t}, \mathbf{X}; \theta_{\text{se}}, \theta_{\text{adaptor}}, \theta_{\text{llm}}) \end{aligned} \quad (1)$$

2.3.1. Stage 1: Adapter-only Within-Language Alignment

Since both the speech encoder and the LLM are pre-trained on large-scale datasets, in the first stage, as Eq. 2, we fine-tune the adaptor using $\mathcal{D}_{\text{mono}}$ to perform an initial coarse-grained alignment. This provides a robust starting point for the subsequent joint optimization of multiple components.

$$\theta_{\text{adaptor}}^* = \arg \min_{\theta_{\text{adaptor}}} \mathcal{L}_{\mathcal{D}_{\text{mono}}} \quad (2)$$

2.3.2. Stage 2: Within-Language Alignment with Progressive Encoder Unfreezing

Following the first-stage fine-tuning of the adaptor, its lightweight design leads to insufficient representational capacity. To achieve more precise alignment in the second stage, as Eq. 3, we progressively unfreeze the speech encoder and optimize it jointly with the adaptor. More specifically, we employ a two-phase unfreezing strategy: first, the last eight layers of the speech encoder are unfrozen and fine-tuned alongside the adaptor; then, the entire speech encoder is activated for full network optimization, and this progressive activation will be discussed in Sec. 4.3. Notably, this stage continues to use only the monolingual dataset.

$$\theta_{\text{adaptor}}^*, \theta_{\text{se}}^* = \arg \min_{\theta_{\text{adaptor}}, \theta_{\text{se}}} \mathcal{L}_{\mathcal{D}_{\text{mono}}} \quad (3)$$

2.3.3. Stage 3: Joint Optimization with LLM-Adaptive

Following the alignment achieved in the first two stages, speech features and textual representations of the corresponding language are now aligned in the semantic space. We then introduce cross-lingual tasks to better leverage the multilingual capabilities of the LLM. However, due to inherent discrepancies in length between speech and text, as well as the rich diversity in speech (such as variations in speaking rate), the granularity of speech representations cannot be strictly matched to that of text. Therefore, in the third stage, as Eq. 4, we unfreeze the LLM and perform joint optimization of the speech encoder, adaptor, and LLM together, enhancing the model’s robustness to such variations.

$$\theta_{\text{adaptor}}^*, \theta_{\text{se}}^*, \theta_{\text{llm}}^* = \arg \min_{\theta_{\text{adaptor}}, \theta_{\text{se}}, \theta_{\text{llm}}} \mathcal{L}_{\mathcal{D}_{\text{mono}} + \mathcal{D}_{\text{cross}}} \quad (4)$$

Through the proposed three-stage progressive alignment framework, our approach effectively bridges the gap between speech and text modalities across languages. This structured training strategy enables the model to fully leverage the LLM’s inherent multilingual capabilities, facilitating robust cross-lingual transfer. As a result, the system achieves enhanced performance in multilingual speech-to-text tasks, offering improved adaptation to variability in speech while maintaining semantic coherence.

3. DATA AND TRAINING SETTING

In this section, we provide a detailed description of the training data sources, test datasets, and evaluation metrics in 3.1. The model configuration and training hyperparameters are specified in 3.2.

3.1. Data and Settings

The training data includes two tasks, ASR (Automatic Speech Recognition) and S2TT (Speech-to-Text Translation). For ASR, there is a total of 810k hours of commercial purchases data covering 10 languages: Chinese (zh), English (en), Japanese (ja), Korean (ko), Cantonese (yue), German (de), French (fr), Russian (ru), Spanish (es), and Italian (it). The S2TT task data encompasses 434k hours, covering language pairs such as zh-en, ja-en, de-en, fr-en, es-en, it-en, ru-en, as well as en-zh, en-ja, en-de, en-sv, en-id, and en-ar. The main sources of S2TT data are: 1) open-source datasets like CoVoST [17], TED-LIUM [18], and MuST-C [19]; 2) constructing S2TT data by translating ASR transcripts into target languages and The rest comes from commercial purchases of 48k hours.

For model capability evaluation, the ASR task includes four types of test sets: LibriSpeech (human-read audiobooks) [20], CommonVoice 15 (crowdsourced speech) [21], Fleurs (human-read Wikipedia) [22], and Wenetspeech (audio from YouTube and podcasts) [23]. CER (Character Error Rate) [24] is used for zh, ja, ko, yue, while WER (Word Error Rate) [25] is used for other languages, all combined with Whisper [26] normalizer post-processing. The evaluation of translation tasks uses primarily CoVoST2 [27], which is based on Common Voice, using BLEU scores for the evaluation, with character-based tokenization for Chinese and Japanese, and the 13a tokenizer for other languages. Details are elaborated in Table 1.

Table 1. Benchmark datasets for ASR and S2TT. “xx→en” means non-English source to English target; “en→xx” means English source to non-English target.

Task	Test Data	Domain	Languages	Metric
ASR	LibriSpeech (test-clean/other)	read	en	WER
	Wenetspeech (test-net/meeting)	YouTube/Podcast	zh	CER
	Fleurs	Wikipedia (read)	zh, en, ja, ko, yue, de, fr, ru, es, it	CER/WER
Common Voice 15			zh, en, ja, ko, yue, de, fr, ru, es, it	CER/WER
S2TT	CoVoST2	crowdsourcing	xx→en: zh, ja, de, fr, es, it, ru en→xx: zh, ja, de, sv, id, ar	BLEU

3.2. Model Specifications

The speech audio encoder is initialized with the SenseVoice-large encoder [28], approximately 700M in size. The LLM is based on Qwen2.5 [29] enhanced with multilingual continuing pre-training in advance. The Adapter consists of two transformer layers and one CNN layer. We experiment with two sizes of the Qwen2.5 model, 1.5B and 7B parameters, corresponding to PART-2B and PART-8B.

Training is performed using 256 NVIDIA A800 GPUs for three epochs, and inference is performed using greedy decoding.

4. EXPERIMENT RESULT

In this section, we conduct a comprehensive set of experiments to evaluate the proposed method. In Sec. 4.1, we compare our approach with several state-of-the-art methods. Sec. 4.2 provides a detailed analysis of the contribution of our two novel designs. Finally, in Sec. 4.3, we perform ablation studies to better understand the role of progressively unfreezing strategy in stage 2.

4.1. Main Result

For Multilingual Automatic Speech Recognition: The experimental results are shown in Table 2. On LibriSpeech, Wenetspeech, Fleurs, and Common Voice 15, PART consistently outperforms the two-stage baseline and remains competitive with larger SLMs. For example, on Fleurs it reduces the average WER from 6.35 to 3.73 ($\downarrow 41\%$), and on Common Voice 15 from 9.18 to 6.29 ($\downarrow 32\%$). These results show that the progressive alignment strategy preserves language-specific features while the task-dependent activation mechanism strengthens cross-language robustness, leading to clear advantages in multilingual ASR.

For Multilingual Speech-to-Text Translation: The experimental results are shown in Table 3. On CoVoST2 across both xx→en and en→xx directions, PART also demonstrates significant advantages. PART-8B achieves the highest BLEU scores on most language pairs, with particularly large gains in low-resource directions such as en→sv, en→id, and en→ar, where it surpasses the two-stage baseline by 3–8 BLEU. This improvement stems from the third-stage introduction of text-based tasks, which, building on modality alignment, further unlocks the LLM’s cross-lingual generation ability, enabling the model to better capture semantics and produce fluent translations in complex multilingual scenarios. At the same time, compared with Whisper-large-v2 and MinMo, PART maintains stable advantages in high-resource languages such as German and French, showing that the method not only addresses cross-lingual alignment challenges but also achieves general improvements in multilingual generation.

4.2. Analysis of Each Stage in Progressive Training

To evaluate the effectiveness of our progressive training strategy, we conducted experiments using mixed ASR and S2TT data under two configurations: two-stage training and three-stage training. The optimization settings and methodologies for both setups were consistent with those described in the Method section. We evaluated ASR performance using Word Error Rate (WER) on the FLEURS test set and S2TT performance using BLEU score on the CoVoST2 test set. As shown in Fig. 2, the model trained with three stages significantly outperformed the two-stage model on both tasks, achieving lower WER and higher BLEU scores. These results underscore the superiority of our progressively fine-grained training strategy.

In addition, we verified that our proposed approach of aligning the model first and then fine-tuning it on cross-lingual tasks contributes to better convergence. As shown in Fig. 2, the three-stage training method uses both ASR and S2TT tasks across all three stages, while PART (Progressive Alignment then Tuning) trains the first two stages only on ASR data and the final stage on both ASR and S2TT tasks. PART achieves lower WER and higher BLEU scores on nearly all subsets. These results confirm that performing

Table 2. ASR main results (WER, %). Bold numbers indicate the best results in each column. “–” means the model does not support the corresponding language or dataset.

Models	Params	LibriSpeech		Wenetspeech		Fleurs							Common Voice 15														
		clean	other	meeting	net	zh	en	ja	ko	yue	de	fr	ru	es	it	zh	en	ja	ko	yue	de	fr	ru	es	it		
SALMON	14B	2.1	4.9	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–		
MinMo	8B	1.7	3.9	6.8	7.4	3.0	3.8	3.8	2.9	4.3	5.2	5.5	6.2	3.4	3.5	6.3	7.9	13.4	6.6	6.4	6.6	8.5	7.0	5.0	6.1	–	
Qwen2-Audio	8B	1.6	3.6	8.1	9.5	7.5	5.1	10.4	10.6	4.1	10.5	9.4	23.2	7.3	6.7	6.9	8.6	13.5	17.5	5.9	7.6	9.6	16.8	5.7	6.8	–	–
Qwen2.5-Omni	8B	1.8	3.4	5.9	7.7	3.0	4.1	–	–	–	–	–	–	–	–	5.2	7.6	–	–	–	–	7.5	–	–	–	–	
PART	2B	2.0	4.2	7.9	7.2	4.2	5.0	3.6	3.1	4.3	5.6	6.5	6.9	3.8	4.2	7.0	9.6	10.2	5.6	5.8	6.4	9.3	7.6	6.8	5.8	–	
PART	8B	1.7	3.8	7.5	6.8	3.9	4.0	3.1	2.5	3.7	4.4	4.7	5.2	2.8	3.0	6.4	8.5	9.7	4.9	5.5	5.1	7.5	5.6	5.1	4.6	–	–

Table 3. S2TT main results (BLEU). Bold numbers indicate the best results in each column. “–” means the model does not support the corresponding language.

Model	Params	xx2en							en2xx						
		zh	ja	de	fr	es	it	ru	zh	ja	de	sv	id	ar	
Whisper-large-v2	1.6B	18.0	26.1	36.3	36.4	40.1	30.9	–	–	–	–	–	–	–	–
Speech-LLaMA	7B	12.3	19.9	27.1	25.2	27.9	25.9	36.8	–	–	–	–	–	–	–
SALMON	14B	–	–	–	–	–	–	–	33.1	22.7	18.6	–	–	–	–
MinMo	8B	26.0	28.9	39.9	41.3	43.3	40.6	48.6	46.7	35.1	–	–	–	–	–
Qwen2-Audio	8B	24.4	20.7	35.2	38.5	40.0	36.3	–	45.2	28.8	29.9	–	–	–	–
Qwen2.5-Omni	8B	29.4	–	37.7	–	–	–	–	41.4	–	30.2	–	–	–	–
LLaST	2B	19.2	24.2	36.8	41.2	43.2	39.3	–	–	–	–	–	–	–	–
PART	2B	23.2	26.8	37.9	39.2	40.9	37.3	46.5	42.4	43.8	30.8	15.7	32.3	16.6	–
PART	8B	27.0	30.0	40.8	42.3	42.7	39.7	50.9	46.8	47.2	35.0	25.8	37.5	22.4	–

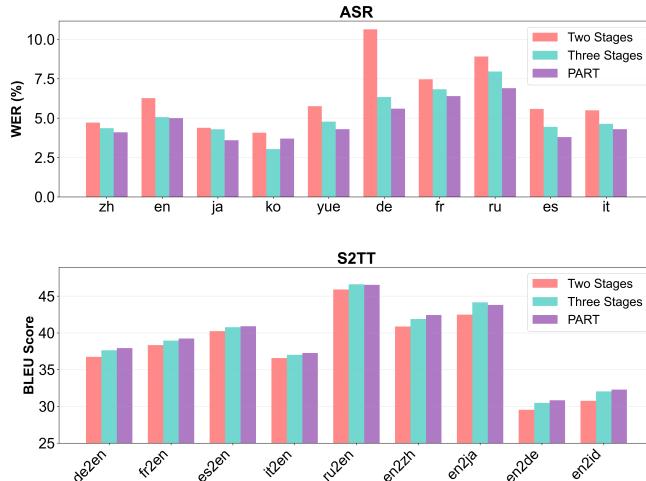


Fig. 2. Comparing WER and BLEU at the progressive training stage alignment before cross-lingual fine-tuning offers a more effective training paradigm for multilingual tasks, as it reduces the language model’s confusion over language-specific speech characteristics.

4.3. Ablation experiments

To isolate and evaluate the impact of progressively unfreezing the speech encoder on both ASR and S2TT tasks, we conducted a controlled ablation study. Both the baseline (“full”) and our progressive method (“last8→full”) were identically trained on the combined dataset $\mathcal{D}_{\text{mimo}} + \mathcal{D}_{\text{cross}}$ following a two-stage protocol: first fine-tuning the adapter modules, followed by joint fine-tuning of the speech encoder and adapter. The sole difference lies in the second stage; the “full” baseline unfreezes and fine-tunes

the entire speech encoder at once, while our “last8→full” strategy progressively unfreezes it (last 8 layers first, then the full encoder). As evidenced in Table 4, our progressive approach yields a consistent improvement, reducing average WER on FLEURS by 0.1 and increasing average BLEU on CoVoST2 by 0.3. These gains confirm that a gradual unfreezing strategy, within this framework, more effectively facilitates alignment learning compared to full-encoder fine-tuning.

Table 4. Ablation study on Stage-2 fine-tuning strategies, comparing full encoder unfreezing against progressive unfreezing (last8→full). Performance is measured by average WER on 10 languages of FLEURS and average BLEU score on 13 language pairs of CoVoST2.

Finetune Paradigm in Stage 2	WER(%)	BLEU
last8- full	6.4	31.7
full	6.3	32.0

5. CONCLUSION

This work proposes PART, a staged, task-dependent training paradigm that decouples intra-language alignment from cross-lingual alignment. The staged approach preserves language specificity while fully exploiting the LLM’s multilingual modeling capacity, avoiding over-convergence of speech representations in multilingual settings. Large-scale evaluations show that PART achieves significant gains on ASR and S2TT. Mechanistic analyses and ablations validate the method’s effectiveness and its marked improvements in robustness to Language Identification (LID) prompts. Overall, PART provides a general and transferable paradigm for multilingual speech–text alignment, and offers valuable insights for future optimization of multilingual speech–language models.

6. REFERENCES

- [1] Jiliang Hu et al., “VHASR: A multimodal speech recognition system with vision hotwords,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, Nov. 2024, pp. 14791–14804, Association for Computational Linguistics.
- [2] Jeongsoo Choi, Se Jin Park, Minsu Kim, and Yong Man Ro, “Av2av: Direct audio-visual speech to audio-visual speech translation with unified audio-visual speech representation,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 27315–27327.
- [3] Tianrui Wang et al., “Viola conditional language models for speech recognition, synthesis, and translation,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 32, pp. 3709–3716, July 2024.
- [4] Minsu Kim et al., “Revival with voice: Multi-modal controllable text-to-speech synthesis,” *arXiv preprint arXiv:2505.18972*, 2025.
- [5] Francesco Verdini et al., “How to connect speech foundation models and large language models? what matters and what does not,” *CoRR*, vol. abs/2409.17044, 2024.
- [6] Marco Gaido et al., “Speech translation with speech foundation models and large language models: What is there and what is missing?”, in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.
- [7] Chao-Wei Huang et al., “Investigating decoder-only large language models for speech-to-text translation,” in *25th Annual Conference of the International Speech Communication Association, Interspeech 2024, Kos, Greece, September 1-5, 2024*, 2024.
- [8] Sai Koneru et al., “Blending LLMs into cascaded speech translation: KIT’s offline speech translation system for IWSLT 2024,” 2024, pp. 183–191.
- [9] Yexing Du et al., “Making LLMs better many-to-many speech-to-text translators with curriculum learning,” in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.
- [10] Yunfei Chu et al., “Qwen2-audio technical report,” *arXiv preprint arXiv:2407.10759*, 2024.
- [11] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al., “Qwen2. 5-omni technical report,” *arXiv preprint arXiv:2503.20215*, 2025.
- [12] Changli Tang et al., “SALMONN: Towards generic hearing abilities for large language models,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [13] Shao-Syuan Huang et al., “Enhancing multilingual asr for unseen languages via language embedding modeling,” in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025.
- [14] Tuan Nguyen, Long-Vu Hoang, and Huy-Dat Tran, “Qwen vs. gemma integration with whisper: A comparative study in multilingual speechllm systems,” *arXiv preprint arXiv:2506.13596*, 2025.
- [15] Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass, “Listen, think, and understand,” in *International Conference on Learning Representations*, 2024.
- [16] Xi Chen et al., “LLaST: Improved end-to-end speech translation system leveraged by large language models,” in *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, Aug. 2024, pp. 6976–6987, Association for Computational Linguistics.
- [17] Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu, “CoVoST: A diverse multilingual speech-to-text translation corpus,” in *Proceedings of The 12th Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 4197–4203, European Language Resources Association.
- [18] Anthony Rousseau et al., “TED-LIUM: an automatic speech recognition dedicated corpus,” in *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 2012.
- [19] Mattia A. Di Gangi et al., “MuST-C: a Multilingual Speech Translation Corpus,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. June 2019, pp. 2012–2017, Association for Computational Linguistics.
- [20] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.
- [21] Rosana Ardila et al., “Common voice: A massively-multilingual speech corpus,” in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 4218–4222, European Language Resources Association.
- [22] Alexis Conneau et al., “Fleurs: Few-shot learning evaluation of universal representations of speech,” in *2022 IEEE Spoken Language Technology Workshop*. IEEE, 2023, pp. 798–805.
- [23] Binbin Zhang et al., “Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022.
- [24] Jesin James, Deepa P Gopinath, et al., “Advocating character error rate for multilingual asr evaluation,” *arXiv preprint arXiv:2410.07400*, 2024.
- [25] Frederick Jelinek, *Statistical methods for speech recognition*, MIT press, 1998.
- [26] Radford et al., “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*. PMLR, 2023, pp. 28492–28518.
- [27] Changhan Wang, Anne Wu, and Juan Pino, “Covost 2 and massively multilingual speech-to-text translation,” *arXiv preprint arXiv:2007.10310*, 2020.
- [28] Keyu An et al., “Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms,” *arXiv preprint arXiv:2407.04051*, 2024.
- [29] Qwen An Yang, Baosong Yang, et al., “Qwen2.5 technical report,” *ArXiv*, vol. abs/2412.15115, 2024.