



Selective Invocation for Multilingual ASR: A Cost-effective Approach Adapting to Speech Recognition Difficulty

Hongfei Xue¹, Yufeng Tang², Jun Zhang², Xuelong Geng¹, Lei Xie^{1*}

¹Audio, Speech and Language Processing Group (ASLP@NPU), School of Software, Northwestern Polytechnical University, China

²ByteDance, China

hfxue@mail.nwpu.edu.cn, lxie@nwpu.edu.cn

Abstract

Although multilingual automatic speech recognition (ASR) systems have significantly advanced, enabling a single model to handle multiple languages, inherent linguistic differences and data imbalances challenge SOTA performance across all languages. While language identification (LID) models can route speech to the appropriate ASR model, they incur high costs from invoking SOTA commercial models and suffer from inaccuracies due to misclassification. To overcome these, we propose SIMA, a selective invocation for multilingual ASR that adapts to the difficulty level of the input speech. Built on a spoken large language model (SLLM), SIMA evaluates whether the input is simple enough for direct transcription or requires the invocation of a SOTA ASR model. Our approach reduces word error rates by 18.7% compared to the SLLM and halves invocation costs compared to LID-based methods. Tests on three datasets show that SIMA is a scalable, cost-effective solution for multilingual ASR applications.

Index Terms: multilingual ASR, selective invocation model, spoken large language models.

1. Introduction

Multilingual automatic speech recognition (ASR) models have gained significant attention for their ability to recognize multiple languages using a single model [1, 2, 3, 4], as illustrated in Figure 1(a). Recent advances have led to impressive performance in various languages through large-scale supervised or self-supervised pre-training [3, 5, 6, 7, 8, 9, 10, 11, 12]. For example, Whisper [6] is trained on 680,000 hours of weakly multilingual data, enabling it to generalize effectively across standard ASR benchmarks, while USM [9] leverages 12 million hours of unlabeled data to achieve robust cross-lingual performance. Despite these advances, the application of multilingual ASR systems with a single model still faces significant challenges. Phonetic differences, syntactic variations, and vocabulary disparities across languages make it difficult to achieve consistent universal state-of-the-art (SOTA) performance. Moreover, imbalances in training data between high-resource and low-resource languages further limit the single-model solutions.

A common strategy to address these challenges is to use a language identification (LID) model that first detects the language of the input speech before invoking the corresponding SOTA ASR model for transcription, as shown in Figure 1(b). However, this two-stage approach has its drawbacks. Many SOTA models are commercial [12] and incur usage fees based on the volume of processing, making this method costly. Additionally, an incorrect LID prediction may trigger the wrong

*Corresponding author.

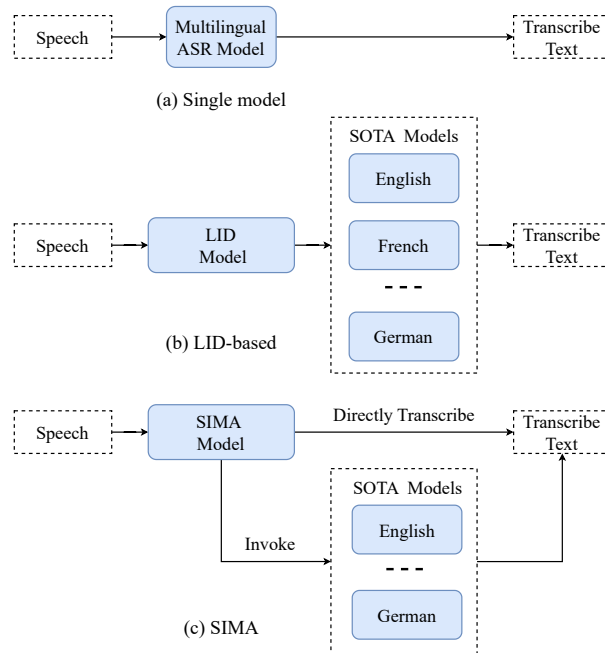


Figure 1: Three systems for multilingual ASR. (a) A single multilingual model, such as Whisper, which recognizes multiple languages with one model. (b) A language identification (LID)-based system that identifies the language and invokes the corresponding SOTA model. (c) Selective invocation for multilingual ASR (SIMA) that directly transcribes simpler speech and invokes SOTA models for more complex inputs.

model, further affecting the user experience [13].

Motivated by these limitations, we propose an alternative strategy that selectively invokes models based on the complexity of the input speech. In ASR tasks, the recognition difficulty varies significantly. Under clean acoustic conditions with simple vocabulary, both the SOTA and regular models typically yield low word error rates (WER). However, in noisy or acoustically challenging environments, the WER increases [14, 15, 16, 17], where robust SOTA models generally perform better [6]. This observation raises a key question: Can we distinguish between simple and complex speech inputs and adapt our ASR system accordingly? Recent advancements in large language models (LLMs) [18, 19, 20] have significantly enhanced the understanding capabilities of spoken large language models (SLLMs) [21, 22, 23, 24, 25]. We hypothesize that SLLMs can not only understand the content of speech but also assess whether they can transcribe the speech accurately. In this way, the SLLM could directly transcribe simple, clean

speech while reserving the invocation of a more robust SOTA model for complex cases.

Based on this hypothesis, we introduce **Selective Invocation** for **Multilingual ASR (SIMA)**, built on a base SLLM. As illustrated in Figure 1(c), SIMA processes the input speech and determines whether to transcribe it directly or invoke a specialized SOTA model for more challenging inputs. Specifically, for simple speech or when the language confidence is low, SIMA generates the transcription itself, thereby avoiding the high costs associated with invoking commercial SOTA models. Furthermore, to more accurately determine whether an invocation is needed, we introduce an “Uncertain” category as a supplement to the binary classification. For these “Uncertain” cases, we employ a fusion confidence strategy that integrates three distinct confidence measures to assess. Experiments on the Multilingual LibriSpeech [26], VoxPopuli [27], and FLEURS [28] datasets demonstrate that SIMA reduces the WER by 18.7% relative to the base SLLM, while halving invocation costs compared to the LID-based approach. These results support our hypothesis and demonstrate that SIMA offers a scalable and cost-effective solution for multilingual ASR applications.

2. Method

As illustrated in Figure 1(c), our proposed method comprises two main modules. The first module is the SIMA model, which evaluates the difficulty of the input speech and either directly produces a transcription or outputs an invocation label. The second module is a library of SOTA ASR models. When the SIMA model yields the invocation label, the input speech is routed to the appropriate SOTA model for transcription.

2.1. SIMA Design

The SIMA model is built upon a multilingual SLLM and is designed to produce one of three possible outputs during training, as depicted in Figure 2. Each output includes a predicted language tag along with an associated language confidence score. The three output types are defined as follows:

Invocation No: For speech deemed sufficiently simple, the model generates transcription tokens followed by an “Invocation No” token, indicating that it can directly transcribe the input. During training, the ground truth transcription and special tokens are used to compute the cross-entropy (CE) loss.

Invocation Yes: For complex speech, the model outputs an “Invocation Yes” token, signaling that a specialized SOTA model should be invoked to process the input. In this case, the CE loss is computed solely based on the special token.

Invocation Uncertain: We found that a binary decision (i.e., simply “Yes” or “No”) can be ambiguous for speech of intermediate difficulty. To address this, we introduce a third category—“Invocation Uncertain”—which indicates uncertainty and the need for further evaluation via a confidence strategy. This output is formatted similarly to the “No” output; however, during training, the transcription tokens are pseudo labels for training the “Transcription Confidence” token, and the CE loss is only computed for special tokens.

It is important to note that when the language confidence score is low, the SIMA model defaults to direct transcription to avoid erroneous invocations.

2.2. Fusion Confidence Strategy

To address cases classified as “Invocation Uncertain” and determine whether a language-specific invocation is necessary, we

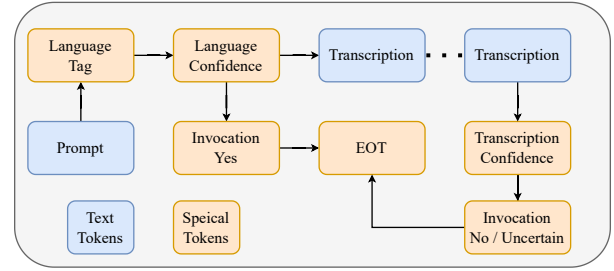


Figure 2: The multitask training format of the SIMA model.

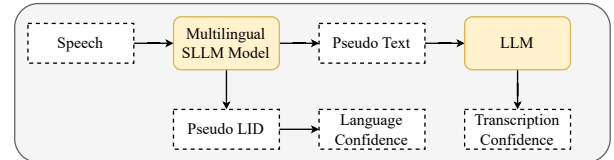


Figure 3: Data pipeline of the SIMA dataset.

integrate three complementary confidence evaluation methods:

Posterior Probability: We extract the maximum probability from the softmax output of the final layer, where $y \in \mathbb{R}^{T \times C}$, with T denoting the number of frames and C representing the number of label classes. The average posterior probability is computed as follows:

$$\text{Probability} = \frac{1}{T} \sum_{t=1}^T \max_{c \in \{1, \dots, C\}} y_{t,c} \quad (1)$$

Entropy: Since entropy is a well-known measure of uncertainty, we employ an entropy-derived confidence score defined by:

$$\text{Entropy} = -\frac{1}{T \times C} \sum_{t=1}^T \sum_{c=1}^C y_{t,c} \log(y_{t,c}) \quad (2)$$

Transcription Confidence: After generating transcription tokens, the SIMA model generates an evaluation score that reflects sentence fluency. This score categorizes the transcription into four confidence levels, ranging from level A (highest) to level D (lowest).

In the fusion strategy, if the posterior probability falls below a threshold P , the entropy exceeds a threshold E , and the transcription confidence is lower than a threshold T , the system determines that the speech recognition is difficult for direct transcription and thus invokes the appropriate SOTA model.

2.3. Data Pipeline

To train the SIMA model, we propose a specialized data pipeline that supports the three distinct output formats. We employ a multilingual SLLM [29] as the base model for generating training data. As illustrated in Figure 3, the SLLM processes the input speech to produce pseudo text labels, from which we compute the corresponding WER. Based on the WER, we assign invocation labels for all languages as follows:

- **Invocation No:** WER in the interval $[0, 2]$.
- **Invocation Yes:** WER exceeding 10.
- **Invocation Uncertain:** WER in the interval $(2, 10]$.

Table 1: WER (%) in LID-based models in MLS testset.

WER	<i>de</i>	<i>en</i>	<i>es</i>	<i>fr</i>	<i>it</i>	<i>nl</i>	<i>pl</i>	Avg
OpenAI	4.68	5.81	4.26	5.82	9.95	9.89	4.52	6.42
Meta	4.83	7.34	4.02	5.45	9.62	12.11	7.00	7.20
Assembly	3.91	5.23	4.73	7.26	13.69	9.81	5.57	7.71
LID-Top	3.91	5.23	4.02	5.45	9.62	9.81	4.52	6.08

Since SOTA model performance (i.e., measured by WER) can vary across languages, we further analyze the results at different WER intervals for each language in Section 3.4.

After obtaining the transcription and invocation labels, we utilize an LLM [30] to score the generated text against the ground truth, thereby producing transcription confidence scores ranging from level A (highest) to level D (lowest). Additionally, the SLLM is used to infer a pseudo LID and its associated posterior probability. If the SLLM misidentifies the language, the transcription is assigned the lowest confidence level (D); otherwise, confidence levels from D to A are determined based on the posterior probability. This design enables the dynamic adjustment of invocation thresholds, reducing the likelihood of erroneous model invocations.

3. Experiments

3.1. Datasets

To ensure domain diversity and improve robustness, we utilize three datasets: Multilingual LibriSpeech (MLS) [26], VoxPopuli [27], and FLEURS [28]. We select the languages common to these datasets, namely English (*en*), German (*de*), Dutch (*nl*), French (*fr*), Spanish (*es*), Italian (*it*), and Polish (*pl*). For training, we generate 100k samples per language using the SIMA data pipeline, with 25k samples sourced from VoxPopuli and 75k from MLS. This division is due to the size of the original dataset. For validation and testing, SIMA data are generated from the original validation and test sets of MLS and VoxPopuli. For FLEURS, only the test set is used as out-of-domain data to evaluate robustness further.

3.2. Experiment Setup

Base Model The base model used for SIMA initialization comprises speech encoders, an adapter, and a text LLM, following the architecture described in IdealLLM [29]. We train the model on the MLS and VoxPopuli datasets to perform both ASR and LID tasks simultaneously, meeting the requirements of SIMA pipeline. It also serves as our baseline for comparison.

SIMA Model The SIMA model is initialized with the base model and fine-tuned on the generated SIMA data using a multi-task learning framework. During training, we maintain a balanced distribution among the three output classes in an approximate ratio of 1:1.5:1.5. For the fusion confidence strategy, we set P to 0.96, E to 0.0015, and T to level B based on the multiple tests on the validation set. Training is performed on 8 NVIDIA A100 GPUs, with gradient accumulation configured to achieve an effective batch size corresponding to 400 seconds of speech per GPU. We use a peak learning rate of $5e-5$, a warmup period of 2k steps, and 10k training steps.

Random Invocation Model To isolate the effectiveness of our proposed method from improvements arising solely from the invocation mechanism, we design a random invocation baseline. In this baseline, the decision to invoke is made randomly at the same overall invocation rate as the SIMA model. Additionally, we propose an **invocation efficiency** metric to quantify the ben-

Table 2: WER (%) ↓ results and Invocation Rate(%) ↓ results.

	MLS		VoxPopuli		FLEURS	
	WER	Rate	WER	Rate	WER	Rate
Whisper	6.42	0	14.00	0	5.06	0
LID-Top	6.08	100	11.01	100	5.06	100
Base	7.86	0	12.43	0	10.76	0
Random	6.85	57.6	11.78	45.5	7.73	51.2
SIMA	6.40	57.6	11.28	45.5	6.43	51.2

Table 3: Others' results of SIMA model and LID-based model.

	MLS	Vox	FLEURS	Avg
SIMA-ACC (%)	72.1	69.7	72.8	71.5
SIMA-F1 (%)	73.6	68.1	68.1	69.9
SIMA-Cost (×)	0.58	0.46	0.51	0.51
SIMA-Invoke-Errors (%)	0.12	0.74	0.18	0.35
LID-Invoke-Errors (%)	0.43	1.22	0.38	0.68

efit of each invocation, defined as the reduction in WER relative to the base model divided by the invocation rate. The invocation rate is the proportion of input speech invoking the SOTA model. **LID-based Model** To establish LID-based benchmarks, we evaluate some open-source models and commercial models on the test sets of MLS, VoxPopuli, and FLEURS. For each dataset, we select the model achieving the lowest WER as the SOTA reference for this study. The models considered for comparison included OpenAI’s API¹, Meta’s SeamlessM4T-large-v2², and AssemblyAI’s universal-1³. For more commercial models, they are not used due to cost constraints. The performance results for MLS are summarized in Table 1, and the same procedure is applied to obtain SOTA results for VoxPopuli and FLEURS. It should be noted that the error caused by the incorrect LID prediction is not considered in LID-Top. Considering the SOTA model uses multilingual models, the impact is not significant.

3.3. Main Results

Table 2 summarizes the performance of the SIMA and baseline models regarding WER and invocation rate across the three test sets. The results indicate that, due to the selective invocation of SOTA models, the SIMA model achieves significant WER reductions of 18.6%, 9.3%, and 28.2% relative to the base model on the three datasets. Furthermore, compared to the random invocation strategy, SIMA consistently delivers lower WER, with improvements of 6.6%, 4.2%, and 16.8%. Notably, the improvement on the FLEURS dataset is especially significant, as it is out-of-domain for the base model but in-domain for the LID-Top model. These findings convincingly demonstrate SIMA’s remarkable ability to precisely determine when to invoke the SOTA model, thereby optimizing overall ASR performance.

Table 3 presents additional performance metrics for SIMA. The invocation decision accuracy (ACC) and F1 scores are approximately 70%, supporting our hypothesis that SLLMs can effectively differentiate speech inputs based on complexity. Although SIMA exhibits a slight WER gap compared to LID-Top, it reduces invocation costs by approximately 0.51× across the three datasets, significantly lowering associated ex-

¹<https://platform.openai.com/docs/guides/speech-to-text>

²<https://huggingface.co/facebook/seamless-m4t-v2-large>

³<https://www.assemblyai.com/research/universal-1>

Table 4: Results of Language-Agnostic and Language-Specific in the MLS test set.

		<i>de</i>	<i>en</i>	<i>nl</i>	<i>es</i>	<i>fr</i>	<i>pl</i>	<i>it</i>	Avg
Language-Agnostic	ACC ↑	72.75	69.01	80.85	76.02	66.24	65.96	73.53	72.05
	F1 ↑	69.01	74.56	88.6	68.61	64.19	71.41	78.71	73.58
	Rate ↓	49.12	72.3	87.71	38.57	53.17	47.31	55.15	57.62
	WER ↓	3.95	5.23	9.65	4.27	5.26	6.29	10.21	6.41
	Efficiency ↑	2.28	1.55	2.60	1.63	0.51	9.07	0.80	2.52
Language-Specific	ACC ↑	72.01	69.83	68.75	72.79	68.88	57.5	73.06	68.97
	F1 ↑	68.58	73.9	70.26	63.47	60.69	61.7	68.35	66.70
	Rate ↓	41.46	67.37	51.32	30.69	41.38	37.31	36.13	43.66
	WER ↓	4.03	5.24	9.87	4.38	5.25	7.04	10.36	6.59
	Efficiency ↑	2.53	1.65	4.01	1.69	0.68	9.49	0.80	2.90

penses. Moreover, incorporating language confidence prediction reduces language invocation errors (SIMA-Invoke-Errors) by about 48.6% relative to LID-based methods, with even greater impact when the SOTA model is monolingual.

3.4. Language-Specific Invocation Strategies

Due to the varying recognition performance (i.e., WER) across different languages in both SOTA and base models, a unified invocation strategy may lead to unnecessary model invocations for some languages. Table 4 compares the unified (language-agnostic) strategy with language-specific strategies. As introduced in Section 2.3, the language-agnostic strategy uses a fixed invocation interval. In contrast, the language-specific strategy customizes this interval: for a given language with a LID-Top model test result of i , the “Uncertain” interval is defined as $(i - 2.5, i + 2.5]$. This tailored approach enhances the efficiency of invoking the SOTA model by adapting to language-specific characteristics.

The results in Table 4 indicate that while language-specific strategies slightly decrease the ACC and F1 scores of invocation decisions (reflecting the increased complexity of the model’s learning task), they also significantly reduce the overall invocation rate, improving the invocation efficiency metric from 2.5 to 2.9. These findings suggest that language-specific invocation strategies can optimize the overall efficiency of the SIMA model by minimizing unnecessary invocations without substantially compromising WER.

3.5. Analysis

Ablation Study Table 5 presents the results of ablation study on MLS, which focuses on two key components: fusion confidence strategy and the use of “Invocation Uncertain”. First, we remove the fusion confidence strategy and replace it with random invocations, forcing all predictions marked as “Uncertain” to be classified definitively as either “Yes” or “No.” This change led to an increased WER, alongside declines in ACC, F1 scores, and efficiency. These findings highlight the confidence strategy’s crucial role in enhancing the precision of model invocations. In the second experiment, we eliminate the “Uncertain” category, compelling the system to make a binary decision. This binary approach increased the overall invocation rate and reduced invocation efficiency, suggesting that it triggers unnecessary invocations for audio samples that are simple and clean. Moreover, the observed decrease in ACC indicates that samples with intermediate difficulty are more susceptible to misclassification when the “Uncertain” option is unavailable.

ASR Performance of SIMA Table 6 compares the ASR per-

Table 5: Ablation study on confidence strategy and uncertainty.

	ACC ↑	F1 ↑	Rate ↓	WER ↓	Efficiency ↑
SIMA	72.1	73.6	57.6	6.40	2.52
– Confidence	69.6	71.5	57.6	6.47	2.24
– Uncertain	67.9	71.3	66.0	6.39	2.21

Table 6: WER (%) results of SIMA’s direct transcription.

	MLS	VoxPopuli	FLEURS	Avg
Base SLLM	7.81	10.79	10.33	9.64
SIMA	7.79	9.72	10.39	9.30

formance of SIMA against the base SLLM. Since SIMA does not generate transcription results for samples that invoke the SOTA model, we evaluate its performance on subsets from the MLS, Vox, and FLEURS datasets—comprising only those samples transcribed directly by SIMA. The results reveal that SIMA training does not increase the WER; in fact, the WER slightly decreases in some cases. This outcome may be attributed to the evaluated subset primarily consisting of simple and clean speech samples that benefit from further training optimization. In contrast, more complex samples (routed to the SOTA model) are excluded from the training loss calculation.

Future Work Although the current SIMA model significantly improves WER, it still lags behind Whisper [6] on out-of-domain data, FLEURS [28]. This limitation stems from our initial hypothesis that the base SLLM model can effectively perform the invoke task. Our base SLLM model [29] is inherently weaker than specialized models such as Whisper because of the limitation of training data. In future work, we plan to adopt Whisper [6] as the base model and further refine the SIMA system to improve the ASR performance of the SOTA model.

4. Conclusion

This paper introduces SIMA, a novel selective invocation strategy for multilingual ASR. Leveraging a base spoken large language model, SIMA dynamically determines whether to transcribe speech directly or invoke specialized SOTA models. Extensive experiments on three benchmark datasets demonstrate that SIMA reduces the word error rate by 18.7% and cuts invocation costs by 51% compared to LID-based methods. These promising results highlight the potential of adaptive ASR systems for scalable, cost-effective real-world applications. In future work, we will explore stronger base models to further enhance performance.

5. References

- [1] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. J. Moreno, E. Weinstein, and K. Rao, "Multilingual speech recognition with a single end-to-end model," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4904–4908.
- [2] A. Conneau and G. Lample, "Cross-lingual language model pre-training," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2019, pp. 7057–7067.
- [3] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2021, pp. 2426–2430.
- [4] J. Shi, D. Berrebbi, W. Chen, E. Hu, W. Huang, H. Chung, X. Chang, S. Li, A. Mohamed, H. Lee, and S. Watanabe, "ML-SUPERB: multilingual speech universal performance benchmark," in *Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2023, pp. 884–888.
- [5] J. Bai, B. Li, Y. Zhang, A. Bapna, N. Siddhartha, K. C. Sim, and T. N. Sainath, "Joint unsupervised and supervised training for multilingual ASR," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6402–6406.
- [6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning (ICML)*, vol. 202, 2023, pp. 28 492–28 518.
- [7] H. Xue, Q. Shao, P. Chen, P. Guo, L. Xie, and J. Liu, "TranUSR: Phoneme-to-word Transcoder Based Unified Speech Representation Learning for Cross-lingual Speech Recognition," in *Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2023.
- [8] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W. Hsu, A. Conneau, and M. Auli, "Scaling speech technology to 1, 000+ languages," *J. Mach. Learn. Res.*, vol. 25, pp. 97:1–97:52, 2024.
- [9] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang *et al.*, "Google usm: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.
- [10] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elshahar, H. Gong, K. Heffernan, J. Hoffman *et al.*, "Seamlessm4t-massively multilingual & multimodal machine translation," *arXiv preprint arXiv:2308.11596*, 2023.
- [11] H. Xue, Q. Shao, K. Huang, P. Chen, J. Liu, and L. Xie, "SSHR: leveraging self-supervised hierarchical representations for multilingual automatic speech recognition," in *International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.
- [12] F. M. Ramirez, L. Chkhetiani, A. Ehrenberg, R. McHardy, R. Botros, Y. Khare, A. Vanzo, T. Peyash, G. Oexle, M. Liang, I. Sklyar, E. Fakhani, A. Etefy, D. McCrystal, S. Flamini, D. Donato, and T. Yoshioka, "Anatomy of industrial scale multilingual ASR," *CoRR*, vol. abs/2404.09841, 2024.
- [13] B. Houston, O. Sadjadi, Z. Hou, S. Vishnubhotla, and K. Han, "Improving multilingual asr robustness to errors in language input," in *Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2024, pp. 1250–1254.
- [14] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2012, pp. 22–25.
- [15] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, vol. abs/1510.08484, 2015.
- [16] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [17] M. Dua, Akanksha, and S. Dua, "Noise robust automatic speech recognition: review and analysis," *Int. J. Speech Technol.*, vol. 26, no. on, pp. 475–519, 2023.
- [18] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.
- [19] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.
- [20] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2308.11276*, 2023.
- [21] J. Wu, Y. Gaur, Z. Chen, L. Zhou, Y. Zhu, T. Wang, J. Li, S. Liu, B. Ren, L. Liu, and Y. Wu, "On decoder-only architecture for speech-to-text and large language model integration," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.
- [22] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.
- [23] S. Hu, L. Zhou, S. Liu, S. Chen, L. Meng, H. Hao, J. Pan, X. Liu, J. Li, S. Sivasankaran, L. Liu, and F. Wei, "Wavlm: Towards robust and adaptive speech large language model," in *EMNLP*. Association for Computational Linguistics, 2024, pp. 4552–4572.
- [24] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.
- [25] H. Xue, Y. Liang, B. Mu, S. Zhang, Q. Chen, and L. Xie, "E-chat: Emotion-sensitive spoken dialogue system with large language models," in *ISCSLP*. IEEE, 2024.
- [26] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," in *Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2020, pp. 2757–2761.
- [27] C. Wang, M. Rivière, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. M. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *ACL/IJCNLP*. Association for Computational Linguistics, 2021, pp. 993–1003.
- [28] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "FLEURS: few-shot learning evaluation of universal representations of speech," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2022, pp. 798–805.
- [29] H. Xue, W. Ren, X. Geng, K. Wei, L. Li, Q. Shao, L. Yang, K. Diao, and L. Xie, "Ideal-LLM: Integrating dual encoders and language-adapted llm for multilingual speech-to-text," *arXiv preprint arXiv:2409.11214*, 2024.
- [30] M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.