LAMA-UT: Language Agnostic Multilingual ASR Through Orthography Unification and Language-Specific Transliteration

Sangmin Lee[1], **Woojin Chung**[1], **Hong-Goo Kang**[1*]

[1]Dept. of Electrical & Electronic Engineering, Yonsei University, South Korea
{sangmin_lee, woojinchung}@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr

Abstract

Building a universal multilingual automatic speech recognition (ASR) model that performs equitably across languages has long been a challenge due to its inherent difficulties. To address this task we introduce a **L**anguage-**A**gnostic **M**ultilingual **A**SR pipeline through orthography **U**nification and language-specific **T**ransliteration (LAMA-UT). LAMA-UT operates without any language-specific modules while matching the performance of state-of-the-art models trained on a minimal amount of data. Our pipeline consists of two key steps. First, we utilize a universal transcription generator to unify orthographic features into Romanized form and capture common phonetic characteristics across diverse languages. Second, we utilize a universal converter to transform these universal transcriptions into language-specific ones. In experiments, we demonstrate the effectiveness of our proposed method leveraging universal transcriptions for massively multilingual ASR. Our pipeline achieves a relative error reduction rate of 45% when compared to Whisper and performs comparably to MMS, despite being trained on only 0.1% of Whisper's training data. Furthermore, our pipeline does not rely on any language-specific modules. However, it performs on par with zero-shot ASR approaches which utilize additional language-specific lexicons and language models. We expect this framework to serve as a cornerstone for flexible multilingual ASR systems that are generalizable even to unseen languages.

Introduction

Developing a model for multilingual automatic speech recognition (ASR) is appealing due to its applicability to universal languages, including low-resource or unseen languages. However, this task presents significant challenges as it requires extensive datasets and involves the complexity of capturing shared characteristics across diverse languages in both phonetic and orthographic domains.

Since the revolution in ASR technologies driven by self-supervised learning (SSL) models (Baevski et al. 2020; Conneau et al. 2020; Hsu et al. 2021), monolingual ASR has achieved superhuman transcription performance, shifting the main focus of recent research towards developing a universal model that spans multiple languages.

There are two primary methods for building a multilingual ASR model. The first approach involves scaling the size of both the labeled dataset and the model itself, using a single universal model to enhance its capacity and cover a vast number (100+) of languages, thereby achieving multilingual ASR (Radford et al. 2023). Another approach involves incorporating language-specific modules into the universal model to address the performance inconsistencies of previous methods. For example, MMS (Pratap et al. 2024) demonstrated the feasibility of scaling multilingual technology to over 1,000 languages by leveraging common features across languages and adding language-specific modules to improve the performance of each language. Indeed, there have been efforts to integrate both methods (Zhang et al. 2023), combining their strengths to build a more robust and versatile model.

Although these works have demonstrated strong performance across various languages, the trade-off between performance and complexity remains a substantial challenge. The first method, using a single universal pipeline, struggles to achieve consistent performance across languages, and its effectiveness in low-resource languages remains uncertain. On the other hand, despite achieving state-of-the-art performance and parameter efficiency, the second method cannot be considered a single universal model due to the inclusion of language-specific modules. Moreover, the use of language-specific modules like adapters, heads, and language models (LMs) sometimes complicates the training and inference pipeline, suggesting potential areas for future improvement.

Simultaneously, large language models (LLMs) have garnered considerable attention for their remarkable capabilities in the natural language processing (NLP) domain. Following this trend, ASR pipelines have integrated audio SSL models as encoders and LLMs as decoders to enhance transcription quality (Li et al. 2023; Fathullah et al. 2024). These approaches involve using projectors (Yu et al. 2024) or fine-tuning strategies (Tang et al. 2024; Du et al. 2024) to align modalities and improve transcription capabilities across multilingual datasets. Subsequently, shallow fusion (Chorowski and Jaitly 2016; Kannan et al. 2018) based scoring methods (Hu et al. 2023; Huang et al. 2024) were attempted to replace conventional LMs with LLMs during the decoding stage. Despite these efforts resulting in per-

*Corresponding author

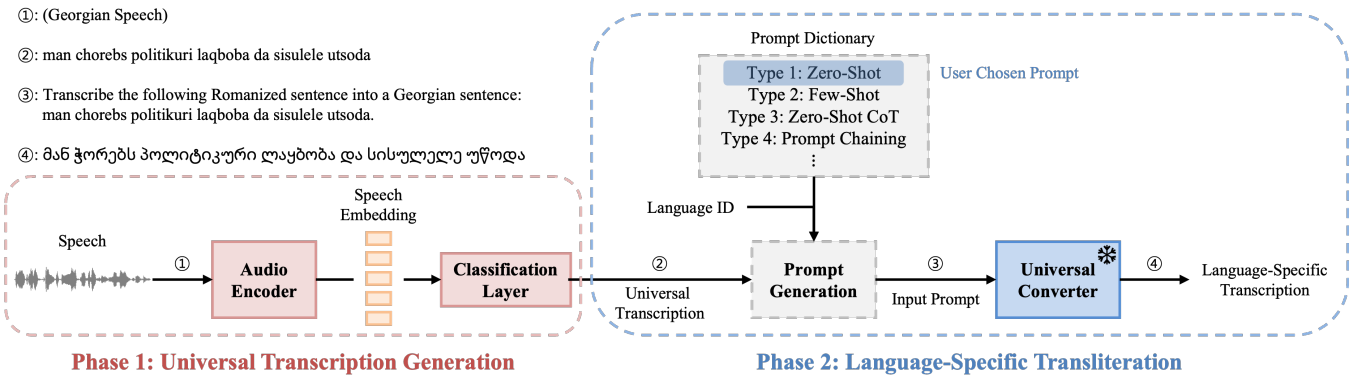

Figure 1: Illustration of our universal ASR pipeline.

formance gain across various languages, a comprehensive method to fully leverage the diverse emergent abilities of LLMs remains to be developed.

In this paper, we introduce a novel language-agnostic multilingual ASR pipeline that spans over 100 languages, including completely unseen languages. As in Fig 1, the proposed pipeline consists of two phases: universal transcription generation and language-specific transliteration. In the universal transcription generation phase, we focused on reducing orthographic complexity by unifying diverse orthographic systems into a consistent format, approximating phonetic features across multiple languages. In the language-specific transliteration phase, we regard the transformation from universal transcription to language-specific transcription as a transliteration task by leveraging a universal converter. Our experiments demonstrate notable transcription performance of LAMA-UT across over 100 languages while using significantly smaller training data (only 680 hours) compared to other state-of-the-art multilingual ASR models. Furthermore, our proposed pipeline outperforms previous methods, especially in low-resource languages, and demonstrates proficiency in completely unseen languages, achieving performance comparable to existing language-agnostic ASR methods without relying on any language-specific modules. Our contributions are summarized as follows:

- We propose a novel language-agnostic multilingual ASR pipeline consisting of two phases: universal transcription generation and language-specific transliteration.
- We enabled our proposed pipeline to perform multilingual ASR with minimal data by unifying diverse orthographic systems through Romanization.
- Our pipeline demonstrates consistent performance across over 100 seen languages and excels with completely unseen languages, all without relying on any language-specific modules or additional fine-tuning.

Related Works

Multilingual ASR

Initially, multilingual ASR models handled a limited number of languages (Toshniwal et al. 2018; Pratap et al. 2020a), under 60. However, recent advancements have led to the development of models capable of managing a broader range of languages. Whisper (Radford et al. 2023) uses a sequence-to-sequence (Sutskever, Vinyals, and Le 2014) approach with 680,000 hours of weakly supervised data, and its neural decoder serves as a LM, enhancing transcription performance. With this method, Whisper attained impressive performance across most supported languages.

Google USM (Zhang et al. 2023) employs a Conformer (Gulati et al. 2020) encoder with various types of heads (Graves 2012; Chan et al. 2016) and is trained on an extensive dataset. It also employs a three-stage training incorporating speech-only, speech-text paired, and text-only data. Furthermore, to enhance transcription performance for low-resource languages, USM integrates language-specific adapters and employs Noisy Student Training (NST) techniques (Xie et al. 2020; Park et al. 2020).

MMS (Pratap et al. 2024), a state-of-the-art multilingual ASR model, employed a Connectionist Temporal Classification (CTC) based approach (Graves et al. 2006) on a dataset covering over 1,000 languages. It utilizes a two-stage fine-tuning pipeline. The first stage involves Romanization-based fine-tuning to learn a global representation across diverse languages. In the second stage, language-specific adapters and heads are added to capture detailed features for each language and fine-tuned.

Zero-Shot ASR

ASR-2K (Li et al. 2022a) is a zero-shot ASR model which utilizes three universal models to cover a range of languages: an acoustic model (Li et al. 2020), a pronunciation model (Li et al. 2022b), and a LM (Scannell 2007). This suggests the potential for a universal multilingual ASR model capable of functioning in a zero-shot environment without relying on any language-specific components. Consequently, Zero-Shot MMS (Zhao, Pratap, and Auli 2024) utilized language-specific lexicon and n-gram LMs in the decoding phase to enhance zero-shot transcription performance.

LLM-Supported Multilingual ASR

Hu et al. 2023 trained a multilingual LLM covering 84 languages and employed a shallow fusion-based per-frame

scoring to enhance transcription quality in multilingual ASR. Subsequently, Huang et al. 2024 introduced non-autoregressive per-segment scoring, which improves transcription performance and reduces the computational burden. These methods primarily leveraged the strengths of a multilingual acoustic model (USM) and achieved further accuracy by incorporating LLMs into the decoding step.

Proposed Method

The overall structure of the proposed multilingual ASR pipeline, LAMA-UT, comprises a universal transcription generation phase and a language-specific transliteration phase, as shown in Fig 1. We produce universal transcriptions by finetuning an audio encoder with an additional classification layer. Consequently, we manually select a prompt type from a predefined dictionary and combine it with language information to generate the input prompt for the universal converter. Finally, by feeding this input prompt into the universal converter, we translate the universal transcription into language-specific ones.

Universal Transcription Generation

Previous studies in linguistics (Ladefoged and Maddieson 1996; Clark, Yallop, and Fletcher 2007) have shown that the phonological characteristics of human speech are constrained by a limited range of sounds due to the anatomical structure of the vocal tract. Similarly, in the ASR domain, prior research (Taguchi and Chiang 2024) has empirically demonstrated that the primary obstacle in multilingual ASR is the orthographic complexity across languages. Through the integration of these two insights, we aim to unify orthographic systems across diverse languages by standardization of notations into a Latin character system. This approach establishes alignment between phonetic and orthographic features through a unified transcription system. As a result, we develop a universal transcription generator capable of producing consistent transcriptions across multilingual speech corpora, including unseen languages.

International Phonetic Alphabet. The first method for orthography unification is to use the international phonetic alphabet (IPA). IPA is a phonetic notation system that includes four elements: consonants, vowels, diacritics, and suprasegmentals. IPA can precisely transcribe pronunciations in a consistent format with a combination of the four elements. There are challenges with the IPA, especially in vocabulary mapping, and one possible solution is to treat the combination of elements as a single token (e.g., ts, dz, etc.). However, due to the vast diversity of possible combinations makes this approach difficult to implement. Conversely, treating each IPA character as a distinct token introduces another issue: characters without phonetic value must be mapped to specific frames as shown in Fig 2. Since diacritics and suprasegmentals provide detailed information about pronunciation (e.g., length, tone, and stress) but do not carry specific phonetic values, mapping them to distinct frames can introduce confusion during the training process.

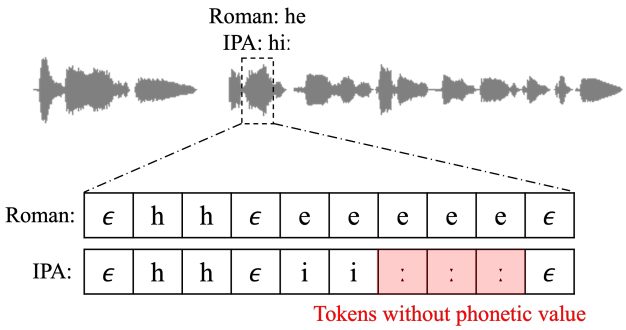

Figure 2: Problems derived in single-token IPA recognition. Diacritic ':' indicates phoneme length, which has no explicit phonetic value. Epsilon denotes a blank token in CTC.

Romanization. Romanization is an alternative method for orthography unification which involves converting text from various writing systems into Latin script. While Romanization does not preserve phonetic features as precisely as the IPA, it generally retains phonetic information. Additionally, Romanization offers several advantages over the IPA. Romanization standardizes diverse writing systems using the Latin alphabet, which is already employed by the majority of languages. In contrast, IPA requires a specific set of rules for converting the orthography of each language into its IPA representation. Thus, Romanization is more efficient as it only requires conversion for languages that do not use the Latin alphabet. Furthermore, Romanization is advantageous for LLMs, as a large portion of their training data consists of Latin characters. Given these benefits, we adopt Romanization as a method for orthography unification.

Universal Transcription Generator. Since our goal is to generate a universal transcription with unified orthography, our first approach was leveraging a wav2vec2.0-phoneme (Xu, Baevski, and Auli 2021). However, we found that directly passing phoneme tokens to the universal converter is suboptimal for transliteration, as it generates phoneme-level tokens without accounting for spacing. To address this, we shifted our focus to developing a universal transcription generator that produces character-level tokens while incorporating spacing information. In this context, Romanization provides a universal character-level orthographic representation, effectively reducing the vocabulary size to around 30 tokens compared to IPA. Since Romanization aligns with the common phonetic features preserved across languages, we are also confident in the proposed method's strong generalization ability for languages not explicitly included in the training data. We selected wav2vec2.0-XLS-R (Babu et al. 2021) with 1 billion parameters, an SSL model pre-trained on 128 languages, as the audio encoder to leverage the advantages of pre-training on a diverse set of languages. We then attach a classification layer on top and finetune both the audio encoder and the classification layer with speech and Romanized transcription pairs to generate universal transcriptions.

Zero-Shot	Transcribe following Romanized sentence into a {lang} sentence: {roman}.
Few-Shot	Here are some examples of transcribing a Romanized sentence into a {lang} sentence: {shots}. Considering the examples above, transcribe the following Romanized sentence into a {lang} sentence: {roman}.
Zero-Shot CoT	Transcribe the following Romanized sentence into a {lang} sentence. Think step by step: {roman}.
Few-Shot + Zero-Shot CoT	Here are some examples of transcribing a Romanized sentence into a {lang} sentence: {shots}. Considering the examples above, transcribe the following Romanized sentence into a {lang} sentence. Think step by step: {roman}.
Prompt Chaining	Transcribe the following Romanized sentence into a {lang} sentence, based on its pronunciation: {roman}.
	Correct the typographical and spacing errors in the following {lang} sentence: {pred}.

Table 1: Specific format of the prompt. *roman* refers to the predicted Romanized transcription, *shots* indicates generated examples sampled from the training data, *pred* denotes output from first prompt and *lang* indicates the name of the language.

Language-Specific Transliteration

The next step is to revert the universal transcription, which retains phonetic features, back to its original language-specific form. Since this process involves a text-to-text transformation, we approach it as a transliteration task. Consequently, we focused on the versatility of LLMs which excel in multilingual and multitask benchmarks due to extensive training on diverse text data. Therefore, we aim to utilize LLMs as universal converters to transform Romanized transcriptions into language-specific ones.

Prompt Generation. While LLMs have brought a tectonic shift to the NLP domain, additional techniques are still needed to fully harness their emergent abilities. In this context, prompt engineering has emerged as a field focused on crafting and refining prompts to effectively utilize LLMs across diverse applications and research areas. To maximize the performance of the inversion process, in the ablation study, we empirically investigated various prompt types: zero-shot, few-shot, zero-shot chain-of-thought (CoT), and prompt chaining, to determine which is the most appropriate for this task.

Universal Converter. We transliterate the unified Romanized transcription by leveraging LLM's multilingual and multitask language understanding ability without finetuning. Since our approach does not require any special finetuning, the universal converter can be replaced with any superior LLMs, potentially improving the performance of our proposed pipeline in line with the rapidly advancing capabilities of LLMs. For this paper implementation, we utilize LLaMA3-8B, 4-bit quantized LLaMA3-70B (Touvron et al. 2023), and GPT-4o-mini (OpenAI 2024) as the universal converter.

Experiments

Dataset

FLEURS. FLEURS (Conneau et al. 2022) is a multilingual speech corpus encompassing 102 languages. It provides a relatively small amount of data per language (approximately 12 hours) while ensuring an unbiased distribution of data across the languages. Given our focus on demonstrating effective multilingual ASR with minimal data, we utilize the FLEURS and its official splits for experiments.

CommonVoice. CommonVoice (Ardila et al. 2020) is a multilingual speech dataset crowdsourced from speakers of various languages. For unseen languages, we leverage the official test split of 25 languages from CommonVoice 17.0, which offers sufficient samples for evaluation.

Data Preprocessing

We initially applied NFKC normalization and lowercase transformation to the text transcriptions. Subsequently, we excluded samples containing parentheses or numbers from the dataset for the following reasons: parentheses and digits in transcriptions introduced ambiguity, as some enclosed phrases were pronounced while others were not, and digits had one-to-many pronunciation mappings across languages (e.g. '1' can be pronounced as 'one', 'eins', 'uno', 'yi', etc.). Finally, we utilized the Python library *Uroman* (Hermjakob, May, and Knight 2018) to obtain Romanized transcription and *Phonemizer* (Bernard and Titeux 2021) for IPA transcription. For Japanese, we employed *Pykakasi* (TAKAHASHI 1992) due to the limitation of *Uroman*, which treats Japanese kanji as Chinese characters. Following these preprocessing steps, we obtained approximately 6 to 8 hours of speech-transcription paired data per language on average.

Training Detail

We performed fine-tuning on all layers except the feature extractor for 3,000 steps with a CTC loss and a batch size of 128. We bypassed the two-stage fine-tuning pipeline from prior studies (Xu, Baevski, and Auli 2021; Pratap et al. 2024) because our distinct methodology, which used a smaller dataset, caused the divided fine-tuning approach to result in premature convergence and instability. For hyperparameters, we employed the default AdamW optimizer (Kingma and Ba 2017; Loshchilov and Hutter 2019) with a tri-stage learning rate scheduler. The warm-up, hold, and decay phases were configured to 10%, 60%, and 30% of the total training steps, respectively. We then performed a series of experiments to determine the optimal learning rate schedule within the range of 5e-6 to 5e-4. Finally, the entire training pipeline was conducted on two RTX-3090 GPUs

Model	Seen (PER / CER ↓)							Unseen (PER / CER ↓)			
	de	nl	fr	es	it	pt	avg	ia	eo	eu	avg
wav2vec2.0-phoneme†	23.8	38.0	31.0	28.7	33.5	45.0	33.0	10.7	-	20.8	31.4
+ n-gram LM†	14.8	26.0	26.4	12.3	21.7	36.5	22.9	**6.1**	-	**13.7**	**22.2**
IPA generator (LAMA-UT)	10.2	10.5	**9.6**	**4.2**	4.6	10.9	14.4	29.0	32.0	36.6	35.1
Roman generator (LAMA-UT)	**7.3**	**9.6**	12.9	4.4	**3.6**	**7.2**	**11.3**	14.0	**20.8**	30.3	32.3

Table 2: Comparison between two orthography unification methods. We report PER and CER for seen and unseen languages. Average values are calculated over all 102 seen and 25 unseen languages, respectively. For a fair comparison, all the model sizes are set to 300 million. † denotes results measured by PER, which does not allow for a strict comparison with other results.

Model	Data (h)	Universal	Zero-Shot
Whisper	680k	O	X
MMS-1162	122k	X	X
LAMA-UT	**0.6k**	**O**	**O**

Table 3: Comparison of previous multi-lingual ASR models with the proposed pipeline. *Data* denotes the total amount of training dataset, *Universal* indicates that no language-specific module is needed, and *Zero-Shot* denotes whether inference on unseen languages is feasible.

	Universal Converter	CER ↓	WER ↓
Seen	LLaMA-8B	26.6	46.7
	LLaMA-70B	15.5	35.3
	GPT-4o-mini	**7.5**	**18.1**
Unseen	LLaMA-8B	33.0	50.2
	LLaMA-70B	27.2	58.9
	GPT-4o-mini	**15.8**	**38.3**

Table 4: Upper bound performance of the universal converter. This upper bound is assessed by feeding ground truth Romanized transcriptions into the universal converter with zero-shot prompting.

with 24GB of VRAM each, and we leveraged gradient accumulation techniques to address memory issues.

Inference Detail

Universal Transcription Generator. We leveraged a beam search decoder from flashlight (Kahn et al. 2022) with a beam size of 100. No additional lexicons or LMs were utilized in the decoding pipeline to maintain a universal pipeline without relying on language-specific elements.

Prompting Strategy. For the prompting strategy, we utilized language information and a subset of the training data to construct our hypothesis prompt. The specific format of the prompt employed is detailed in Table 1. In zero-shot prompting, the universal converter automatically transforms Romanized transcriptions into language-specific ones using only the Romanized transcriptions and language information. We employed zero-shot prompting to evaluate the performance of the LLM with minimal input. Few-shot prompting (Brown et al. 2020) involves providing examples to help the model generate responses to subsequent instances. We hypothesized that this approach would be particularly effective for low-resource or unseen languages by inducing in-context learning. Specifically, we randomly sampled five Romanized transcription and target transcription pairs for each few-shot example. Zero-shot CoT prompting (Kojima et al. 2022) is a technique that supports complex reasoning by inducing the decomposition of intricate tasks into detailed steps. Specifically, we appended the phrase "Let's think step by step" to the input prompt to encourage the reasoning of the model. Prompt chaining employs a sequence of prompts, with each prompt building upon the output of the previous one, to manage complex multi-step tasks. In this aspect, we concentrated on the decomposable process of converting predicted Romanized transcriptions into language-specific transcriptions through *(i) reverse-Romanization* and *(ii) error correction*. We considered that errors in Romanized transcriptions could propagate during transliteration to language-specific ones, potentially reducing system performance.

Universal Converter. Finally, we required the output of the universal converter to conform to a specific format. We instruct the model to enclose the output within three backticks (e.g., ```), which allows us to isolate and sort only the language-specific transcription from the output of the model. We set the temperature value to 0.0 for all LLMs to obtain deterministic results.

Results

Table 3 shows that LAMA-UT effectively achieves multilingual ASR with a universal model. This approach even operates in a zero-shot environment without requiring language-specific modules while utilizing only a minimal amount of training data. In the subsequent results, we aim to validate the performance of each component within the pipeline.

Universal Transcription Generator

Comparison to Baseline Model. We conducted a performance comparison between our universal transcription generator and the existing baseline, wav2vec2.0-phoneme (Xu, Baevski, and Auli 2021). Our universal transcription generator focuses on generating character-level tokens and is measured using Character Error Rate (CER), while the baseline wav2vec2.0-phoneme is measured using Phoneme Er-

Resource	Lang.	Whisper-large-v3			MMS-1162			LAMA-UT		
		Data (h)	CER ↓	WER ↓	Data (h)	CER ↓	WER ↓	Data (h)	CER ↓	WER ↓
High	es	11000	1.2	3.1	2969	1.6	5.8	**6.1**	2.8	7.3
	it	2585	0.5	1.6	1566	1.2	5.2	**6.8**	2.0	5.2
	id	1014	1.4	5.7	71	2.9	14.2	**6.8**	4.2	11.5
Middle	ta	136	18.3	26.7	265	11.0	41.5	**6.3**	19.5	31.9
	ur	104	30.9	65.0	57	9.0	29.0	**4.9**	14.9	31.9
	sk	90	2.9	8.7	301	2.2	8.8	**4.5**	3.8	10.2
Low	mk	16	10.3	26.3	45	1.5	8.1	**5.1**	5.5	17.2
	hi	12	35.9	43.3	57	5.8	19.6	**5.0**	8.2	15.0
	kk	12	8.5	35.1	46	2.8	15.2	**8.1**	6.7	22.9
Average		-	23.9	42.9	-	**7.8**	**28.8**	-	**14.8**	**33.2**

Table 5: Comparison results with the baseline models. Average CER and WER have reported over 82 languages from FLEURS that are covered by Whisper, MMS, and our method. The classification of the amount of resources is based on the volume of training data used by Whisper. We utilized MMS which encompasses 1162 languages, trained on a combined dataset from MMS-lab, FLEURS, CommonVoice, Voxpopuli (Wang et al. 2021), and MLS (Pratap et al. 2020b).

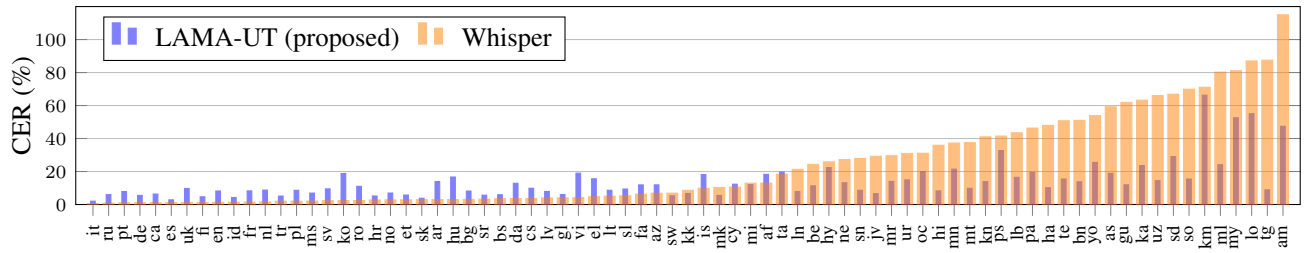

Figure 3: CER comparison between LAMA-UT and Whisper

ror Rate (PER). However, since both metrics are fundamentally used for estimating phonetic symbols, this comparison can be considered meaningful. Following the Table 2, the results show that the proposed method demonstrated significantly better performance across a broader range of languages compared to existing approaches even not utilizing language-specific modules (e.g. n-gram LM). Furthermore, our pipeline demonstrated relatively strong transcription capabilities for unseen languages that were not explicitly included in the training data. In conclusion, transcribing diverse languages based on their pronunciation can produce a universal transcription which is highly effective.

Orthography Unification Methods. Among the two methods for standardizing orthographic features, Romanization proved to be more effective than IPA. Its ability to represent pronunciation across languages while reducing complexity makes it a superior choice for meaningful results. Romanization balances phonetic accuracy with simplicity, providing better alignment with LLMs and ensuring efficient processing across multilingual ASR tasks. However, since these results are constrained to the first phase, we have constructed the end-to-end performance comparison between IPA-based and Romanization-based LAMA-UT, and the results are shown in the appendices[1].

[1]https://github.com/sanghyang00/LAMA-UT-Appendices

Universal Converter Verification

Despite the effectiveness of orthography unification, the success of the entire pipeline hinges on the proper functioning of the universal converter. Therefore, the most critical aspect to validate before experimentation was whether a frozen LLM could effectively serve as a universal converter. To validate this objective, we passed ground truth Romanized transcriptions into the frozen universal converter and assessed its performance. This approach not only tests the converter's capability to accurately produce language-specific transcriptions but also serves to evaluate the upper bound performance of the universal converter within the proposed ASR pipeline. In Table 4, results demonstrated that universal transcription based on pronunciation characteristics can yield significant performance improvements compared to previous works when the universal transcription generator operates ideally. However, the upper bound performance for unseen languages showed a slight decrease compared to seen languages. This decrease is likely because the unseen languages we tested are typically extremely low-resource languages within the training data of the LLM.

Overall Pipeline

Seen Languages. We leveraged two baseline models for comparison: Whisper and MMS. In Table 5, results demonstrate that the proposed method achieved a relative reduction

Model	Repetition Rate (%) ↓	Prompting Strategy									
		Zero-Shot		Few-Shot (5)		Zero-Shot CoT		Few-Shot (5) + Zero-Shot CoT		Prompt Chaining	
		CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓
LLaMA-8B	12	35.1	70.6	22.7	49.6	37.2	77.4	22.1	49.8	35.9	70.8
LLaMA-70B	1	24.3	53.8	17.4	43.8	25.4	54.6	16.8	43.7	26.7	55.2
GPT-4o-mini	**0.2**	**16.6**	**39.3**	**15.3**	**37.2**	18.2	41.0	15.7	37.9	16.9	38.7

Table 6: Effects of prompting strategy and model type on the universal converter. The repetition rate indicates the proportion of samples with format errors (e.g., no section enclosed in three backticks until the maximum token limit) due to word repetition.

Model	Data (h)	# Lang.	Universal	CER ↓
ASR-2K	2k	8	O	65.5
LAMA-UT (Roman)	**0.6k**	**102**	**O**	**34.7**
MMS-ZS	40k	1078	X	29.2
+ n-gram LM	40k	1078	X	25.2

Table 7: Comparison with previous zero-shot approaches. We evaluated transcription quality on 25 unseen languages from the CommonVoice 17.0 dataset. *# Lang.* denotes the number of languages leveraged in training.

of 60% in CER and 30% in WER compared to Whisper. Moreover, LAMA-UT matches the performance of MMS despite the absence of language-specific adapters, heads, and n-gram LMs. Notably, the performance improvements were most pronounced for low-resource languages. While Whisper exhibited increased error rates for these languages due to limited training data, our method showed substantial performance enhancements with minimal data resources. Despite the slight performance degradation in high-resource languages, the improvement observed in low-resource languages is remarkably meaningful. The full comparison results are presented in Fig 3. It is noteworthy that these results were achieved with considerably smaller training data compared to Whisper and MMS.

Unseen Languages. Our main focus was developing a generalized pipeline that demonstrates strong performance with unseen languages. To validate this objective, we utilized two zero-shot ASR models as baselines: ASR-2K and Zero-Shot MMS (MMS-ZS). In Table 7, our method demonstrated a reduction in CER by half while using significantly less training data compared to ASR-2K. Furthermore, it is noteworthy that our proposed pipeline performs remarkably well even without language-specific modules, demonstrating comparable performance to MMS-ZS which leverages language-specific lexicon and n-gram LM.

Ablation Study

Prompting Strategy. In Table 6, few-shot prompting showed the highest performance across all models and prompting strategies. Interestingly, even with zero-shot prompting, the proposed pipeline consistently outperforms Whisper on average in CER and WER, where Whisper records 23.9% and 42.9% respectively, as shown in Table 5. On the other hand, the use of sequential reasoning failed to achieve the anticipated improvements. Specifically, we observed considerable error propagation when utilizing zero-shot CoT prompts and prompt chaining techniques. Minor inaccuracies in the Romanization phase were amplified as they were processed by the LLM, leading to transcriptions that deviated in meaning from the intended output.

Model Size and Training Data. From the perspective of model size, using a relatively smaller LLM like LLaMA-8B frequently resulted in issues such as word repetition, which complicated the transcription sorting process. Additionally, this model faced challenges with language misprediction, often generating transcriptions in languages other than the intended target language. This issue was particularly noticeable with low-resource languages such as Arabic. With the LLaMA-70B model, while word repetition was less pronounced compared to the LLaMA-8B model, the issue of language misprediction persisted, albeit at a reduced frequency. Among the LLMs tested, GPT-4o-mini demonstrated the best performance overall. It outperformed the other models across all prompting strategies, achieving an impressive average CER of 15% across 102 languages.

Conclusion

In this paper, we introduced a generalized multilingual ASR pipeline LAMA-UT, that operates effectively without relying on language-specific modules. By utilizing Romanized transcription as a unified representation across languages, we structured the multilingual ASR pipeline into two phases. Initially, Romanization aligns phonetic and orthographic features, allowing the universal transcription to be effectively generalized across diverse languages and trained efficiently with a smaller dataset. Subsequently, we used a frozen LLM to convert the universal transcription into language-specific ones. This inversion process showed remarkable performance across languages, including those not previously encountered. Our experiments demonstrated that the proposed method not only maintains performance for high-resource languages but also significantly outperforms existing methods for low-resource languages, all while effectively handling unseen languages. Furthermore, our approach matched the performance of models that employ language-specific modules, despite not using any such components. We anticipate that this research will provide a viable alternative for utilizing LLMs to support universal multilingual ASR systems across a variety of applications.

References

Ardila, R.; Branson, M.; Davis, K.; Kohler, M.; Meyer, J.; Henretty, M.; Morais, R.; Saunders, L.; Tyers, F.; and Weber, G. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4218–4222. Marseille, France: European Language Resources Association.

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. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. arXiv:2111.09296.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, 12449–12460. Curran Associates, Inc.

Bernard, M.; and Titeux, H. 2021. Phonemizer: Text to Phones Transcription for Multiple Languages in Python. *Journal of Open Source Software*, 6(68): 3958.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Chan, W.; Jaitly, N.; Le, Q.; and Vinyals, O. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4960–4964. IEEE.

Chorowski, J.; and Jaitly, N. 2016. Towards better decoding and language model integration in sequence to sequence models. arXiv:1612.02695.

Clark, J.; Yallop, C.; and Fletcher, J. 2007. *An Introduction to Phonetics and Phonology*. Blackwell Textbooks in Linguistics. Wiley. ISBN 9781405130837.

Conneau, A.; Baevski, A.; Collobert, R.; Mohamed, A.; and Auli, M. 2020. Unsupervised Cross-lingual Representation Learning for Speech Recognition. arXiv:2006.13979.

Conneau, A.; Ma, M.; Khanuja, S.; Zhang, Y.; Axelrod, V.; Dalmia, S.; Riesa, J.; Rivera, C.; and Bapna, A. 2022. FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech. arXiv:2205.12446.

Du, Z.; Wang, J.; Chen, Q.; Chu, Y.; Gao, Z.; Li, Z.; Hu, K.; Zhou, X.; Xu, J.; Ma, Z.; Wang, W.; Zheng, S.; Zhou, C.; Yan, Z.; and Zhang, S. 2024. LauraGPT: Listen, Attend, Understand, and Regenerate Audio with GPT. arXiv:2310.04673.

Fathullah, Y.; Wu, C.; Lakomkin, E.; Jia, J.; Shangguan, Y.; Li, K.; Guo, J.; Xiong, W.; Mahadeokar, J.; Kalinli, O.; et al. 2024. Prompting large language models with speech recognition abilities. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 13351–13355. IEEE.

Graves, A. 2012. Sequence Transduction with Recurrent Neural Networks. arXiv:1211.3711.

Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, 369–376.

Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; and Pang, R. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Interspeech 2020*, 5036–5040.

Hermjakob, U.; May, J.; and Knight, K. 2018. Out-of-the-box Universal Romanization Tool uroman. In *Proceedings of ACL 2018, System Demonstrations*, 13–18. Melbourne, Australia: Association for Computational Linguistics.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29: 3451–3460.

Hu, K.; Sainath, T. N.; Li, B.; Du, N.; Huang, Y.; Dai, A. M.; Zhang, Y.; Cabrera, R.; Chen, Z.; and Strohman, T. 2023. Massively multilingual shallow fusion with large language models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Huang, W. R.; Allauzen, C.; Chen, T.; Gupta, K.; Hu, K.; Qin, J.; Zhang, Y.; Wang, Y.; Chang, S.-Y.; and Sainath, T. N. 2024. Multilingual and fully non-autoregressive asr with large language model fusion: A comprehensive study. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 13306–13310. IEEE.

Kahn, J. D.; Pratap, V.; Likhomanenko, T.; Xu, Q.; Hannun, A.; Cai, J.; Tomasello, P.; Lee, A.; Grave, E.; Avidov, G.; et al. 2022. Flashlight: Enabling innovation in tools for machine learning. In *International Conference on Machine Learning*, 10557–10574. PMLR.

Kannan, A.; Wu, Y.; Nguyen, P.; Sainath, T. N.; Chen, Z.; and Prabhavalkar, R. 2018. An analysis of incorporating an external language model into a sequence-to-sequence model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5828. IEEE.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*, volume 35, 22199–22213. Curran Associates, Inc.

Ladefoged, P.; and Maddieson, I. 1996. *The sounds of the world's languages*, volume 1012. Blackwell Oxford.

Li, X.; Dalmia, S.; Li, J.; Lee, M.; Littell, P.; Yao, J.; Anastasopoulos, A.; Mortensen, D. R.; Neubig, G.; Black, A. W.; et al. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8249–8253. IEEE.

Li, X.; Metze, F.; Mortensen, D. R.; Black, A. W.; and Watanabe, S. 2022a. ASR2K: Speech Recognition for Around 2000 Languages without Audio. In *Proc. Interspeech 2022*, 4885–4889.

Li, X.; Metze, F.; Mortensen, D. R.; Watanabe, S.; and Black, A. W. 2022b. Zero-shot learning for grapheme to phoneme conversion with language ensemble. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2106–2115.

Li, Y.; Wu, Y.; Li, J.; and Liu, S. 2023. Prompting large language models for zero-shot domain adaptation in speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1–8. IEEE.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Park, D. S.; Zhang, Y.; Jia, Y.; Han, W.; Chiu, C.-C.; Li, B.; Wu, Y.; and Le, Q. V. 2020. Improved Noisy Student Training for Automatic Speech Recognition. In *Interspeech 2020*. ISCA.

Pratap, V.; Sriram, A.; Tomasello, P.; Hannun, A.; Liptchinsky, V.; Synnaeve, G.; and Collobert, R. 2020a. Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters. arXiv:2007.03001.

Pratap, V.; Tjandra, A.; Shi, B.; Tomasello, P.; Babu, A.; Kundu, S.; Elkahky, A.; Ni, Z.; Vyas, A.; Fazel-Zarandi, M.; Baevski, A.; Adi, Y.; Zhang, X.; Hsu, W.-N.; Conneau, A.; and Auli, M. 2024. Scaling Speech Technology to 1,000+ Languages. *Journal of Machine Learning Research*, 25(97): 1–52.

Pratap, V.; Xu, Q.; Sriram, A.; Synnaeve, G.; and Collobert, R. 2020b. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Interspeech 2020*. ISCA.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Scannell, K. 2007. The Crúbadán Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora (WAC3-2007): Proceedings of the 3rd Web as Corpus Workshop, Incorporating Cleaneval*, volume 4, 5. Presses univ. de Louvain.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Taguchi, C.; and Chiang, D. 2024. Language Complexity and Speech Recognition Accuracy: Orthographic Complexity Hurts, Phonological Complexity Doesn't. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15493–15503. Bangkok, Thailand: Association for Computational Linguistics.

TAKAHASHI, H. 1992. KAKASI-Simple Kana Kanji Converter. *http://kakasi. namazu. org/*.

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*.

Toshniwal, S.; Sainath, T. N.; Weiss, R. J.; Li, B.; Moreno, P.; Weinstein, E.; and Rao, K. 2018. Multilingual speech recognition with a single end-to-end model. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4904–4908. IEEE.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Wang, C.; Riviere, M.; Lee, A.; Wu, A.; Talnikar, C.; Haziza, D.; Williamson, M.; Pino, J.; and Dupoux, E. 2021. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 993–1003. Online: Association for Computational Linguistics.

Xie, Q.; Luong, M.-T.; Hovy, E.; and Le, Q. V. 2020. Self-Training With Noisy Student Improves ImageNet Classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Xu, Q.; Baevski, A.; and Auli, M. 2021. Simple and Effective Zero-shot Cross-lingual Phoneme Recognition. arXiv:2109.11680.

Yu, W.; Tang, C.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; Ma, Z.; and Zhang, C. 2024. Connecting speech encoder and large language model for asr. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12637–12641. IEEE.

Zhang, Y.; Han, W.; Qin, J.; Wang, Y.; Bapna, A.; Chen, Z.; Chen, N.; Li, B.; Axelrod, V.; Wang, G.; Meng, Z.; Hu, K.; Rosenberg, A.; Prabhavalkar, R.; Park, D. S.; Haghani, P.; Riesa, J.; Perng, G.; Soltau, H.; Strohman, T.; Ramabhadran, B.; Sainath, T.; Moreno, P.; Chiu, C.-C.; Schalkwyk, J.; Beaufays, F.; and Wu, Y. 2023. Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages. arXiv:2303.01037.

Zhao, J.; Pratap, V.; and Auli, M. 2024. Scaling A Simple Approach to Zero-Shot Speech Recognition. arXiv:2407.17852.