MoLGE: Mixture of Language Group Experts for Efficient Scaling of Massively Multilingual Speech Recognition

Sangmin Lee, Woojin Chung, Woongjib Choi, and Hong-Goo Kang
Dept. of Electronics and Electrical Engineering, Yonsei University, Seoul, South Korea
{sangmin_lee,woojinchung,woongzip1}@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr

Abstract

Massively multilingual automatic speech recognition (ASR) models covering hundreds of languages must maintain robust performance across diverse linguistic and acoustic conditions. However, these models often encounter the ***curse of multilinguality***, where model capacity is diluted across languages. To address this challenge, we propose Mixture of Language Group Experts (MoLGE), built upon speech self-supervised models (S3Ms). MoLGE assigns dedicated expert modules to clusters of similar languages, reducing the number of required submodules compared to conventional language-specific Mixture-of-Experts (MoE) schemes. It further integrates a hierarchical Low-Rank Adaptation (LoRA) strategy into the disentangled acoustic and linguistic components of the S3M architecture, enabling efficient modeling of language-specific characteristics while maintaining parameter efficiency. Further, we investigate the impact of language grouping strategies based on both linguistic and data-driven criteria on overall performance, providing an interpretable perspective on how language structure influences scalability in multilingual speech systems. In experiments, we evaluate MoLGE on a multilingual benchmark encompassing 495 languages. Results demonstrate that MoLGE consistently outperforms dense multilingual baselines with a minimal increase in trainable parameters. Notably, these language grouping strategies yield substantial improvements for both phonetic and orthographic aspects of ASR modeling. Our findings suggest that structured language specialization provides an effective pathway for massively scaling language coverage of multilingual ASR.

1 Introduction

Scaling automatic speech recognition (ASR) to hundreds of languages within a unified model remains a core challenge in multilingual speech processing. While these systems aim for broad accessibility, they must navigate significant linguistic and acoustic diversity. Consequently, performance often degrades as the number of supported languages increases, a phenomenon primarily driven by limited model capacity relative to the task's complexity.

Since the advent of speech self-supervised models (S3Ms) (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022), speech representation learning has advanced significantly by leveraging large-scale unlabeled corpora. These models have achieved robust performance across various downstream tasks, including ASR, in some cases surpassing human-level transcription accuracy on monolingual benchmarks. Consequently, recent research has increasingly focused on extending S3Ms to multilingual settings (Conneau et al., 2021; Babu et al., 2022; Zanon Boito et al., 2024; Pratap et al., 2024; Chen et al., 2024). Current approaches primarily rely on scaling training data to millions of hours (Zhang et al., 2023) and increasing model capacity to billions of parameters (Keren et al., 2025) to improve multilingual coverage and cross-lingual generalization.

Despite these advances, significant challenges remain. Increasing dataset size and model capacity is computationally demanding and often impractical at scale (Zhang et al., 2024;

Yang et al., 2025). Recent massively multilingual systems already require hundreds of millions to billions of parameters, rendering both pretraining and adaptation resource-intensive, particularly in computationally constrained settings (Lugo & Vielzeuf, 2024).

Furthermore, multilingual speech models are typically trained on highly imbalanced datasets dominated by high-resource languages. This imbalance exacerbates the curse of multilinguality (Conneau et al., 2020; Gurgurov et al., 2024), as fixed model capacity is shared across many languages while being disproportionately optimized for a small subset. Consequently, a single dense multilingual model often underperforms relative to specialized monolingual systems (Wang et al., 2020; Chen et al., 2025), highlighting the limitations of uniform scaling for multilingual speech modeling. These limitations reveal a fundamental tension between language scalability and resource accessibility. While scaling multilingual models aims to support underrepresented languages, their increasing computational demands driven by model size and training data limit their practical utility in the settings they are intended to serve (Zhong et al., 2024; Cahyawijaya, 2024; Pava et al., 2025).

To this end, we argue that this tension can be alleviated by shifting from uniform capacity scaling to structured architectural design informed by language relationships. Languages are not independent entities; they share phonological, morphological, and genealogical properties that define a structured space of linguistic similarity (Bloomfield, 1983; Littell et al., 2017). Exploiting this structure as an architectural prior allows model capacity to be allocated in a linguistically principled manner, providing a more efficient pathway for multilingual scaling than exhaustive parameter or data expansion.

Building on this motivation, we propose ***Mixture of Language Group Experts (MoLGE)***, a parameter-efficient and language-aware framework for massively multilingual speech modeling. MoLGE enables structured language specialization by incorporating linguistic relationships directly into its expert allocation mechanism, without relying solely on scaling.

Specifically, our framework is built on two design principles. First, motivated by evidence that lower S3M layers capture universal acoustic features while higher layers encode language-specific information (Pasad et al., 2021; 2023), we adopt a speech-oriented adaptation strategy. This includes hierarchical adaptation, where shared LoRA modules model universal acoustics in lower layers and language group-specific experts capture linguistic characteristics in higher layers, as well as shared experts and an attentive pooling router to improve acoustic modeling and stabilize expert specialization.

Second, we explore language grouping as a mechanism for injecting linguistic prior knowledge into the model architecture. We hypothesize that linguistically or acoustically similar languages share structural patterns, and that grouping them under shared experts effectively increases the accessible training signal for each expert, acting as a form of implicit data augmentation. To this end, we investigate six grouping strategies spanning implicit, embedding-guided, and knowledge-guided approaches, providing a comprehensive evaluation of how different types of linguistic prior influence expert specialization.

Beyond the architectural design, we conduct a systematic analysis of when and why language-aware modeling is most beneficial. By evaluating MoLGE across two ASR models with distinct output characteristics, phonetic (language-agnostic Romanized transcription) and orthographic (language-specific scripts), we examine how the complexity of the target linguistic space modulates the effectiveness of structured language grouping.

Our contributions are summarized as follows:

- We propose MoLGE, an efficient and scalable framework for massively multilingual ASR that incorporates language structure as an architectural prior for expert specialization.
- We integrate speech modeling principles into LoRA and MoLE, incorporating hierarchical adaptation, shared experts, and an attentive pooling router to improve acoustic modeling and expert specialization.
- We provide a systematic study of six language grouping strategies across 495 languages, showing that explicit linguistic priors consistently outperform implicit or random priors, and that their benefit scales with the complexity of the target linguistic space.

2 Related Work

2.1 Multilingual Speech Foundation Models

Multilingual speech foundation models have primarily advanced by scaling model capacity and training data to support an increasing number of languages. Early research, such as XLSR-53 (Conneau et al., 2021), demonstrated that multilingual self-supervised pretraining with wav2vec 2.0 (Baevski et al., 2020) improves cross-lingual transfer and generalization for ASR. Subsequent models, including XLS-R (Babu et al., 2022), MMS (Pratap et al., 2024), mHuBERT-147 (Zanon Boito et al., 2024), AfriHuBERT (Alabi et al., 2025), XEUS (Chen et al., 2024), and Omnilingual ASR (Keren et al., 2025), have extended this paradigm by scaling to hundreds or thousands of languages and increasing model capacity to the billion-parameter regime. In parallel, large-scale supervised Whisper-style approaches (Radford et al., 2023; Peng et al., 2025) have achieved robust multilingual ASR performance by training on massive labeled datasets. Despite these advances, existing methods predominantly rely on uniform scaling of model size and training data, with limited attention to language-aware modeling that explicitly accounts for linguistic diversity.

2.2 Scalable Parameter-Efficient Finetuning

Parameter-efficient fine-tuning (PEFT) methods have been widely adopted to adapt large pretrained models without updating the entire parameter set. Approaches such as adapters (Houlsby et al., 2019; Pfeiffer et al., 2021) and Low-Rank Adaptation (LoRA) (Hu et al., 2022) enable efficient task-specific adaptation by introducing a minimal number of trainable parameters. These methods have been successfully applied across both natural language processing (NLP) and speech domains to maintain performance while reducing computational overhead. In multilingual speech modeling, PEFT has been used to scale models across numerous languages. For example, Pratap et al. (2024) employs language-specific adapters to extend dense multilingual models to thousands of languages. However, assigning individual adapters to every language becomes impractical as the scale increases.

To improve scalability, Mixture-of-Experts (MoE) (Jacobs et al., 1991) approaches introduce conditional computation by activating only a subset of parameters for each input. Recent research further integrates MoE with PEFT, such as Mixture of LoRA Experts (MoLE) (Wu et al., 2024), enabling expert specialization through lightweight adaptation modules. Nevertheless, existing MoE-based multilingual methods typically allocate experts at the granularity of individual languages (Gaur et al., 2021; Kwon & Chung, 2023; Mu et al., 2025), which limits their scalability in massively multilingual scenarios.

Recent studies in NLP have begun exploring language grouping within MoE frameworks (Zheng et al., 2024; Li et al., 2025a; Liang et al., 2025; Janeiro et al., 2025), suggesting that structured parameter sharing across related languages improves scalability. However, such approaches remain underexplored in massively multilingual speech modeling. Existing methods typically apply language-specific LoRA modules to only a small subset of languages (Wang et al., 2023; Bagat et al., 2025; Li et al., 2025b) or focus on specialized scenarios such as code-switching (Ma et al., 2023; Huang et al., 2025), limiting their applicability to large and diverse language sets.

2.3 Position of Our Work

Our work departs from prior approaches in two key dimensions.

First, instead of relying on uniform scaling, we integrate PEFT with speech-aware adaptation and language-group-based routing to enable structured, scalable specialization for a massive set of languages. Second, while prior group-aware routing has primarily focused on textual inputs, we address the unique challenges of language grouping from a speech-oriented perspective by conducting an in-depth analysis of various language grouping strategies.

Taken together, we present an efficient scaling method for massively multilingual speech recognition, along with the first systematic study of language grouping and its impact.

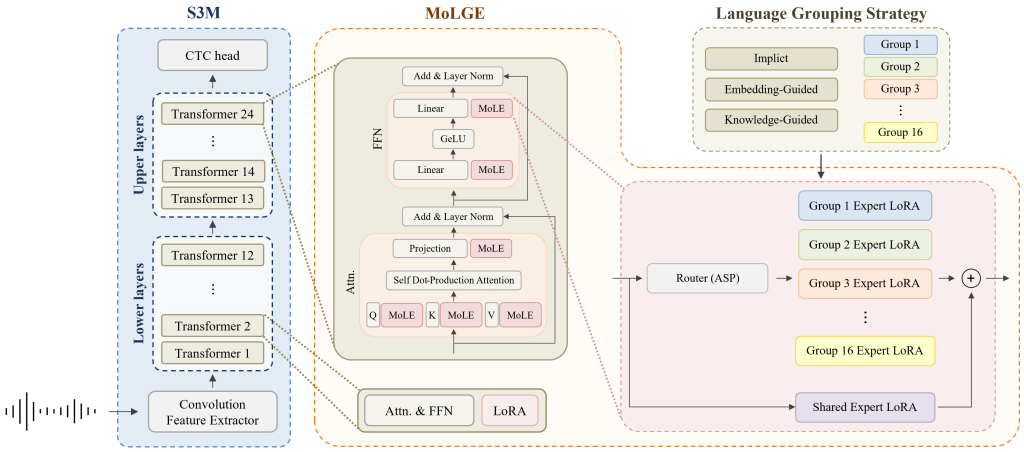


Figure 1: Overall architecture of MoLGE. The conventional S3M is decomposed into lower acoustic and upper linguistic layers, each with distinct adaptation strategies.

3 Proposed Method

We investigate the role of language grouping in multilingual speech recognition through two primary hypotheses. First, grouping linguistically or acoustically similar languages enables experts to capture common structural characteristics, improving parameter efficiency and representation quality, particularly in massively multilingual settings. Second, under a fixed model capacity (i.e., a constrained number of experts), explicit language grouping strategies based on linguistic or statistical priors facilitate more effective expert specialization than implicit, router-driven grouping learned solely from data.

3.1 Mixture of Language Group Experts

In this section, we present ***Mixture of Language Group Experts (MoLGE)***, a speech-centric MoLE framework designed for efficient expert specialization through structured language grouping. In contrast to conventional approaches that allocate experts at the individual language level, MoLGE clusters related languages and assigns shared experts to each group. This strategy significantly improves scalability in massively multilingual settings. The overall architecture of MoLGE is illustrated in Figure 1.

We adopt a hierarchical modeling approach based on the layer-wise functional specialization of speech encoders, where lower layers primarily encode language-agnostic acoustic features and higher layers capture language-dependent linguistic representations (Pasad et al., 2021; 2023). Following this observation, we apply *shared LoRA* modules to the lower layers to model global acoustic variability. Since acoustic characteristics across human speech are largely transferable, a unified parameter space efficiently captures cross-lingual acoustic factors without introducing routing overhead. Conversely, we deploy *Mixture of LoRA Experts (MoLE)* exclusively in the upper layers. This design allocates expert capacity to the linguistic representation space, where language-specific disambiguation is most critical, enabling efficient modeling of diverse multilingual environments.

To further adapt the MoLE framework for speech recognition, we introduce two key modifications. Unlike text, speech signals encode both linguistic content and paralinguistic information, such as speaking rate and speaker identity. We introduce a *shared expert* to capture these paralinguistic factors, which remain largely invariant across different languages. This allows the language-specific experts to focus more effectively on distinct linguistic representations. In addition, we employ an *Attentive Statistical Pooling (ASP) classifier* (Okabe et al., 2018) as the routing mechanism. Given that speech signals often contain temporally imbalanced information, such as long silences, ASP utilizes attention-weighted temporal pooling to emphasize salient acoustic regions while suppressing irrelevant frames. This ensures the router identifies more informative representations for expert assignment.

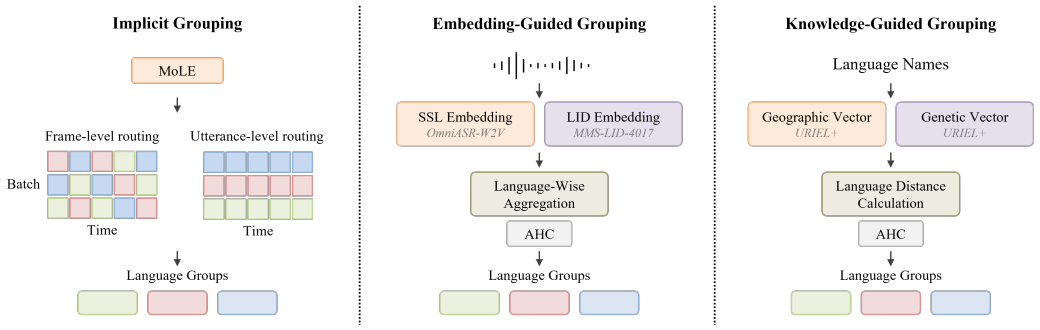


Figure 2: Illustration of the overall procedure for language grouping, covering implicit, embedding-guided, and knowledge-guided strategies.

3.2 Language Grouping Strategy

In this section, we construct six language grouping strategies based on diverse information sources. We consider three categories: *implicit*, *embedding-guided*, and *knowledge-guided*, each comprising two variants. The detailed grouping procedure is illustrated in Figure 2.

Following the principles of statistical cluster analysis (Milligan & Cooper, 1985), the optimal number of clusters can be approximated as $k \approx \sqrt{N/2}$, where N denotes the number of data points. Given our experimental setting of approximately 500 languages ($N \approx 500$), this heuristic suggests $k \approx 15.8$. Accordingly, we fix the number of experts to 16 across all strategies. This choice provides a practical trade-off between model capacity and computational efficiency in massively multilingual scenarios.

Implicit Grouping. Implicit grouping refers to a strategy in which languages are not assigned to experts via predefined labels; instead, groupings emerge naturally from the learned behavior of the router. This represents the foundational form of MoE and has been widely studied in the literature (Lepikhin et al., 2021; Fedus et al., 2022; Jiang et al., 2024).

Under this strategy, we evaluate two distinct routing granularities: frame-level and utterance-level routing (Fan et al., 2024). Frame-level routing treats each timestep of the speech embedding as an individual token, operating analogously to token-level routing in MoE models in NLP. It allows for fine-grained, dynamic expert assignment throughout the duration of a signal. In contrast, utterance-level routing treats an entire utterance as a single unit, routing the complete sequence to a specific expert via a shared router. This ensures consistency in expert application across the temporal dimension of the speech input.

Embedding-Guided Grouping. Embedding-guided grouping derives language clusters from high-dimensional representations generated by pretrained speech foundation models. Under this strategy, we first extract speech embeddings and construct language prototype vectors by aggregating these representations at the language level. Then, we apply Agglomerative Hierarchical Clustering (AHC) with Ward's linkage (Ward Jr, 1963) to partition the languages. This clustering approach is motivated by the inherently hierarchical and phylogenetic structure of human languages (Bloomfield, 1983; François, 2015; Heggarty et al., 2010; Kalyan & François, 2018). Moreover, Ward's method minimizes within-cluster variance to encourage the formation of balanced and compact clusters. This structural property is critical for ensuring stable load balancing and mitigating collapse during training.

To evaluate the impact of different representational biases, we implement this strategy using two distinct pretrained speech models. The first variant utilizes OmniASR-W2V [1], a multilingual wav2vec2.0-style model trained on 1,239 languages. Because this model is optimized with a general self-supervised learning (SSL) objective, it captures diverse speech characteristics, including speaker, phonetic, and language-level information. The second variant employs MMS-LID-4017 [2], a language identification model which trained on and is

[1] https://huggingface.co/facebook/omniASR-W2V-1B
[2] https://huggingface.co/facebook/mms-lid-4017

capable of classifying 4,017 languages. In contrast to the pure SSL model, MMS-LID-4017 is specifically optimized for language identification (LID), thereby emphasizing features that distinguish one language from another. By comparing these two architectures, we evaluate two routing paradigms: SSL-embedding- and LID–embedding–based routing, respectively.

Knowledge-Guided Grouping. Knowledge-guided grouping constructs language clusters by leveraging established linguistic priors. To provide a principled and quantitative representation of this knowledge, we utilize *URIEL+* language vectors (Littell et al., 2017; Khan et al., 2025). These vectors integrate curated linguistic databases into language-level representations, covering 7000+ languages. It also incorporates associated distance metrics derived from a multidimensional attribute space encompassing geographical proximity, typological features, morphological structures, and script-related properties.

Specifically, we construct language distance matrices using two distinct categories of URIEL+ vectors: *geographic* and *genetic*. Geographic vectors capture spatial proximity between languages, capturing regional contact and shared acoustic environments. In contrast, genetic vectors encode genealogical relationships through language family and subfamily hierarchies. This design facilitates typology-aware grouping grounded in shared linguistic ancestry and historical evolution rather than surface-level feature similarity. Consistent with the embedding-guided approach, final language clusters are obtained by applying AHC with Ward's linkage to the resulting distance matrices.

3.3 Two-Stage Dense-to-Sparse Upcycling

We apply a two-stage upcycling strategy to the S3M-based ASR models, inspired by LLM upcycling approaches that transform dense models into sparse MoE architectures (Komatsuzaki et al., 2023; He et al., 2024; Lee et al., 2024). This approach allows us to preserve the rich representations learned by dense models while enabling efficient scaling, reducing training cost, and improving parameter utilization in sparse settings. In the first stage, a dense ASR model is trained on top of pretrained S3M to establish robust global representations and transcription capabilities across languages. In the second stage, we introduce the MoLGE framework to enable expert specialization across predefined language groups, upcycling the dense model into a sparse, parameter-efficient architecture while further improving performance. During the second stage, we apply LoRA to all linear layers, including attention projections and feed-forward networks (FFNs). Motivated by prior work suggesting that FFNs encode model knowledge (Yun et al., 2020; Geva et al., 2021; Dai et al., 2022) while attention layers control information aggregation (Abnar & Zuidema, 2020; Ferrando et al., 2022), we adapt both components. Restricting adaptation to attention layers alone is insufficient for modeling high-density speech representations, making FFN adaptation essential for stable performance across diverse languages.

4 Experiments

4.1 Datasets

We combine three multilingual ASR datasets for training—FLEURS (Conneau et al., 2023) (968 hours, 102 languages), CommonVoice (Ardila et al., 2020) (10,448 hours, 131 languages), and the Omnilingual ASR Corpus (Keren et al., 2025) (2,342 hours, 348 languages)—totaling 13,758 hours across 495 unique languages. All samples were resampled to 16 kHz. As the Omnilingual ASR corpus mainly consists of long utterances, we segment them into approximately 30-second chunks using MMS-FA (Pratap et al., 2024). The segmented OmniASR dataset is available at HuggingFace[3], and additional details on preprocessing procedures are provided in Appendix A.

[3]https://huggingface.co/datasets/Sanghyang00/omniasr-molge

Low-Resource: <10h, **Mid-Resource**: 10–50h, **High-Resource**: >50h

Backbone	Method	Activated Params.	Average		Low-Resource		Mid-Resource		High-Resource	
			CER (↓)	Acc. (↑)	CER (↓)	Acc. (↑)	CER (↓)	Acc. (↑)	CER (↓)	Acc. (↑)
LAMA-UT (Phonetic)	Dense	300 M	25.15	-	26.16	-	19.06	-	21.84	-
	Random	343 M	24.31	68.44	25.46	69.88	17.66	77.77	19.93	23.84
	MoLGE	**343 M**	**22.27**	**93.96**	**23.43**	**94.84**	**15.55**	**93.43**	**17.99**	**80.51**
OmniASR (Orthographic)	Dense	300 M	29.30	-	31.41	-	20.68	-	13.38	-
	Random	353 M	25.81	53.24	27.59	53.32	18.12	67.84	13.37	19.37
	MoLGE	**353 M**	**23.73**	**94.44**	**25.44**	**95.27**	**16.34**	**89.35**	**11.72**	**91.91**

Table 1: Performance on LAMA-UT and Omnilingual ASR (OmniASR) models. We report CER (%) and routing accuracy (Acc., %) breakdown by resource availability.

4.2 Backbone and Baseline Models

We evaluate the MoLGE framework using two multilingual ASR models as a backbone, which emphasize complementary aspects of speech recognition: ***phonetic*** and ***orthographic*** modeling. For a fair comparison, both models utilize Connectionist Temporal Classification (CTC) variants for ASR and are configured with 300M parameters. Details on the rationale behind the backbone selection are provided in Appendix B.

LAMA-UT. LAMA-UT (Lee et al., 2025) is a multilingual ASR framework that predicts a universal Romanized transcription in a language-agnostic manner, subsequently transliterating it into language-specific scripts using frozen LLMs. In our experiments, we use only the Romanized transcription generator to isolate the model's ***phonetic*** modeling capabilities.

Omnilingual ASR. Omnilingual ASR (Keren et al., 2025) is a massively multilingual ASR model based on wav2vec2.0, supporting over 1600 languages. In contrast to LAMA-UT, this model directly predicts language-specific scripts, emphasizing ***orthographic*** modeling.

Baseline Configurations. For each backbone, we compare MoLGE against two baselines: (1) a *dense model*, which uses a single ASR model without any adaptation, corresponding to the original CTC-based ASR architecture, and (2) *random grouping*, where MoLGE is applied with randomly assigned language groups without any linguistic guidance.

4.3 Training Details

We configured the MoLGE with a fixed LoRA rank of 32 and trained it with a composite objective of CTC loss (Graves, 2012) and a routing loss, where implicit grouping uses load balancing and explicit grouping uses cross-entropy for supervised group prediction. The routing loss is weighted by 0.01 relative to the CTC loss. All models are trained with AdamW (Loshchilov & Hutter, 2019) optimizer using a tri-stage learning rate schedule over 10k steps: linear warm-up from $5e^{-6}$ to $1e^{-4}$ (10%), constant (60%), and decay to $5e^{-6}$ (30%).

To alleviate resource and environmental imbalance in multilingual data, we adopt a multi-level sampling strategy following Pratap et al. (2024) and Keren et al. (2025), formulated as $p_{l,d} \propto (n_{l,d}/N)^{\beta_{L,D}}$. Here, $n_{l,d}$ denotes the number of samples for language l (or corpus d), and N is the total number of samples. We set $\beta_L = \beta_D = 0.5$. Additionally, we also apply data augmentation using room impulse responses (RIR) (Szöke et al., 2019) and background noise (Thiemann et al., 2013), where 50% of samples are clean, and the rest are equally split between RIR convolution and additive noise (25% each).

5 Results and Analysis

5.1 Effects of MoLGE on Massively Multilingual ASR

Performance Across Linguistic Domains. As shown in Table 1, applying MoLGE yields substantial improvements in both phonetic and orthographic modeling for speech recognition, with relative gains of 11.5% and 19.0% in CER, respectively. Interestingly, improvements in the phonetic domain are relatively modest compared to the orthographic domain, and

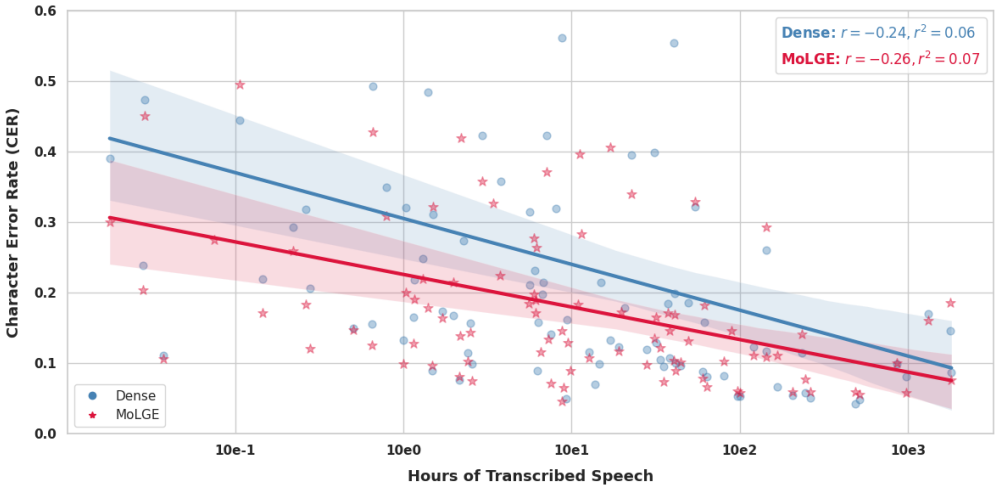
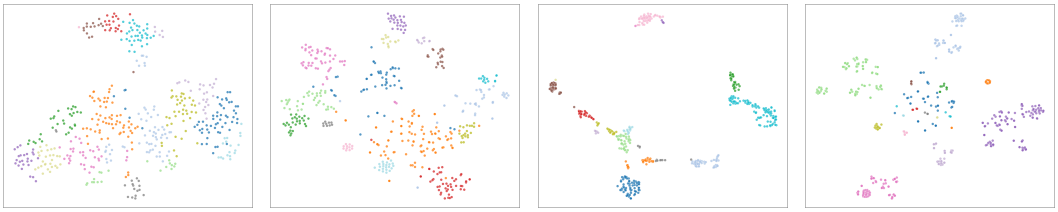


Figure 3: Regression plot between CER and training data. Each point indicates the mean CER per language, and the shaded region represents the confidence interval.



(a) SSL (b) LID (c) Geographic (d) Genetic

Figure 4: Visualization of language clustering across different grouping strategies. Each point represents a language, and colors indicate cluster assignments in the t-SNE space.

mid-resource languages sometimes exhibit lower CER than high-resource languages. This suggests that performance in the phonetic domain (i.e., a simplified latent space (Jaavid et al., 2024; Saji et al., 2025; Liu et al., 2025)) is influenced not only by data scale but also by factors such as representation complexity. In contrast, in the orthographic domain, MoLGE significantly improves both CER and routing accuracy compared to dense and randomly grouped baselines, highlighting the benefits of incorporating language-aware structure in more complex linguistic spaces. Overall, these results indicate that the effectiveness of MoLGE is closely tied to linguistic complexity, with linguistic priors becoming increasingly beneficial for modeling richer orthographic and language-specific structures.

Data–Performance Correlation and Generalization Ability. In Figure 3, we illustrate the regression plot, which visualizes the correlation between CER and training data in the orthographic domain (i.e., OmniASR). When applying MoLGE, the correlation between training data size and CER becomes marginally stronger, with the correlation coefficient r decreasing from -0.24 to -0.26, leading to more predictable learning dynamics. Moreover, the regression slope becomes less steep, with CER on low-resource languages dropping substantially from 42% to 30%, demonstrating that MoLGE narrows the performance gap across resource levels. Furthermore, confidence intervals become notably narrower, indicating that languages with similar training data now achieve more consistent performance. Taken together, these results demonstrate that MoLGE enhances robustness in low-resource languages and enables reliable generalization across languages of varying resource levels.

LAMA-UT (Phonetic)				OmniASR (Orthographic)			
Guidance	Method	CER (↓)	Acc. (↑)	Guidance	Method	CER (↓)	Acc. (↑)
Implicit	Frame	22.30	-	Implicit	Frame	24.30	-
	Sample	22.32	-		Sample	24.51	-
Embedding	SSL	22.29	72.17	Embedding	SSL	23.82	64.16
	LID	**22.27**	**93.96**		**LID**	23.74	**94.70**
Knowledge	**Geographic**	22.29	93.51	Knowledge	**Geographic**	**23.73**	94.44
	Genetic	22.29	92.24		Genetic	23.80	90.79

Table 2: Average transcription performance (CER) and routing accuracy (Acc.) on six guidance methods, respectively.

S.E.: Shared Expert
H.A.: Hierarchical Adapt.

Model	CER (↓)	Acc. (↑)
MoLGE	**23.73**	**94.44**
− S. E.	24.70	92.77
− H. A. (M)	23.85	73.12
− H. A. (L)	27.24	-

Table 3: Ablation study on architecture design.

Indic Languages			Nigerian Languages		
Language	Hours	CER (↓)	Language	Hours	CER (↓)
Tamil	89.13	14.73	Hausa	12.24	12.58
Bengali	40.38	10.68	Yoruba	10.82	22.46
Marathi	11.44	9.90	Tangale	7.01	20.15
Chitwania Tharu	5.43	10.70	Burak	6.70	19.82
Odia	5.05	19.01	Tula	5.68	21.63
Santali (Ol Chiki)	0.07	21.21	Cakfem-Mushere	5.53	24.00

Table 4: ASR performance of MoLGE on six Indic and Nigerian languages, respectively.

# Experts	Total Params.	CER (↓)
8	409M	27.34
16	465M	**24.51**
32	580M	24.52

Table 5: Effect of the number of experts in sample-level implicit grouping.

5.2 Analysis on Cluster Formation and Language Grouping

Cluster Formation Across Guidance Methods. Figure 4 illustrates clustering results using t-SNE across four explicit grouping strategies. Embedding-guided methods generally produce weakly separated clusters, with SSL representations showing particularly fragmented structures. While LID-based clustering yields clearer boundaries than SSL, it remains more dispersed than knowledge-guided methods, likely due to the high-dimensional nature of embeddings (e.g., a feature dimensionality of 1,280 for both SSL and LID models).

In contrast, knowledge-guided methods produce more structured and well-separated clusters than embedding-guided approaches, owing to the compact nature of knowledge vectors. Geographic grouping forms coarse but coherent clusters, while genetic grouping leaves some languages scattered near the center. We attribute this behavior to the sensitivity of knowledge-guided methods to the imputation mechanism of URIEL+. Since URIEL+ imputes missing features based on information from related or higher-level language families, genetic grouping often assigns linguistically diverse languages—including isolates such as Japanese (jpn) and Korean (kor), dialects such as Liberian English (lir), and Maria (mrr), and indigenous languages such as Hunjara-Kaina Ke (hkk) and Ömie (aom)—to the same group. Additional visualizations with language annotations are provided in Appendix D.

Effect of Cluster Coherence. Table 2 presents the performance of MoLGE under various grouping configurations. Across both phonetic and orthographic spaces, explicit guidance consistently outperforms implicit guidance, supporting our hypothesis that prior-based grouping is more effective than router-driven grouping.

To further analyze this trend, we break down the results by representation space. In the phonetic space, performance differences across grouping strategies are minimal, also suggesting that language priors play a limited role in a simplified phonetic latent space. In contrast, in the orthographic space, incorporating language priors consistently improves performance, and more coherent grouping leads to lower CER, as evidenced by the superior performance of LID over SSL embeddings and geographic over genetic vectors.

Overall, these findings indicate that the effectiveness of language grouping arises from the interaction between linguistic complexity, the information density of linguistic priors, and cluster coherence. As linguistic complexity increases, richer priors yield more coherent groupings, which in turn translate into improved performance.

5.3 Knowledge Transfer Across Resource Levels

Table 4 presents ASR performance on representative mid- and low-resource languages from India and Nigeria. To investigate whether MoLGE promotes knowledge sharing across resource levels, we compare training hours and recognition performance within each language group.

As shown in Table 4, languages belonging to the same linguistic family are commonly assigned to the same group and exhibit relatively consistent CER trends despite large differences in training hours. For example, several low-resource languages achieve CERs comparable to those of substantially higher-resource languages within the same group. Similar patterns are observed across both the Indic and Nigerian language groups. Overall, these findings suggest that MoLGE facilitates knowledge transfer within each language group, helping reduce performance disparities between high- and low-resource languages.

5.4 Ablation Study

Architectural Design. In Table 3, we conduct an ablation study by sequentially removing key components, starting with the shared experts in the upper Transformer layer MoLE module, followed by the speech-aware hierarchical adaptation. Removing the shared experts degrades both CER and routing accuracy, highlighting their importance in effective acoustic modeling within the MoLE framework.

Furthermore, eliminating the hierarchical adaptation layer consistently deteriorated model performance. Replacing MoLGE with pure MoLE (indicated by 'M') slightly impaired performance despite higher parameter counts, while replacing it with vanilla LoRA (indicated by 'L') led to a severe performance drop. This confirms the necessity of hierarchical adaptation for integrating language-specific acoustic and textual features.

Number of Experts. In addition to the theoretical motivation discussed in Section 3.2, we conducted a pilot study to examine the impact of the number of experts. As shown in Table 5, reducing the number of experts from 16 to 8 resulted in a substantial degradation in recognition performance. Increasing the number of experts to 32, however, provided negligible improvement (CER from 24.51 to 24.52) while increasing the model size from 465M to 580M parameters. Based on these empirical and theoretical bases, we adopted 16 experts, which offers the best trade-off between recognition performance and computational efficiency.

6 Conclusion

In this paper, we introduced Mixture of Language Group Experts (MoLGE), an efficient and scalable framework for massively multilingual automatic speech recognition. We show that incorporating architectural inductive biases on linguistic structure, together with language grouping, improves transcription performance while enabling scalable modeling across a large number of languages and mitigating the curse of multilinguality. Specifically, our approach integrates speech-oriented parameter-efficient finetuning with structured language grouping, providing a unified framework for scalable massively multilingual speech recognition. We further conduct a comprehensive analysis of diverse grouping strategies, demonstrating their effectiveness across varying resource conditions and their role in promoting expert specialization. Finally, experiments across phonetic and orthographic domains reveal that the benefits of language-aware modeling become more pronounced as linguistic complexity increases, with explicit language priors providing more coherent and effective guidance. We hope this work serves as a foundation for scaling multilingual speech systems and inspires future research on structured, language-aware modeling.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government Ministry of Science and ICT (MSIT) (RS-2026-25468664).

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in Transformers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020.

Jesujoba O. Alabi, Xuechen Liu, Dietrich Klakow, and Junichi Yamagishi. AfriHuBERT: A self-supervised speech representation model for African languages. In *Interspeech 2025*, pp. 4023–4027, 2025.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 4218–4222, 2020.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, pp. 2278–2282, 2022.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460. Curran Associates, Inc., 2020.

Raphaël Bagat, Irina Illina, and Emmanuel Vincent. Mixture of LoRA experts for low-resourced multi-accent automatic speech recognition. In *Interspeech 2025*, pp. 1143–1147, 2025.

Mathieu Bernard and Hadrien Titeux. Phonemizer: Text to phones transcription for multiple languages in python. *Journal of Open Source Software*, 6(68):3958, 2021.

Shikhar Bharadwaj, Chin-Jou Li, Yoonjae Kim, Kwanghee Choi, Eunjung Yeo, Ryan Soh-Eun Shim, Hanyu Zhou, Brendon Boldt, Karen Rosero Jacome, Kalvin Chang, et al. Prism: Benchmarking phone realization in speech models. *arXiv preprint arXiv:2601.14046*, 2026.

Leonard Bloomfield. An introduction to the study of language. 1983.

Samuel Cahyawijaya. *LLM for everyone: Representing the underrepresented in large language models*. Hong Kong University of Science and Technology (Hong Kong), 2024.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

William Chen, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiatong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu, and Shinji Watanabe. Towards robust speech representation learning for thousands of languages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 10205–10224. Association for Computational Linguistics, 2024.

William Chen, Jinchuan Tian, Yifan Peng, Brian Yan, Chao-Han Huck Yang, and Shinji Watanabe. OWLS: Scaling laws for multilingual speech recognition and translation models. In *International Conference on Machine Learning*, 2025.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, 2020.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021*, pp. 2426–2430, 2021.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: Few-shot learning evaluation of universal representations of speech. In *IEEE Spoken Language Technology Workshop*, pp. 798–805. IEEE, 2023.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained Transformers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, 2022.

Dongyang Fan, Bettina Messmer, and Martin Jaggi. Towards an empirical understanding of MoE design choices. In *ICLR Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Javier Ferrando, Gerard I Gállego, and Marta R Costa-Jussà. Measuring the mixing of contextual information in the Transformer. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 8698–8714, 2022.

Alexandre François. Trees, waves and linkages: Models of language diversification. In *The Routledge Handbook of Historical Linguistics*, pp. 161–189. Routledge, 2015.

Neeraj Gaur, Brian Farris, Parisa Haghani, Isabel Leal, Pedro J Moreno, Manasa Prasad, Bhuvana Ramabhadran, and Yun Zhu. Mixture of informed experts for multilingual speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6234–6238. IEEE, 2021.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.

Alex Graves. Connectionist temporal classification. In *Supervised Sequence Labelling with Recurrent Neural Networks*, pp. 61–93. Springer, 2012.

Daniil Gurgurov, Tanja Bäumel, and Tatiana Anikina. Multilingual large language models and curse of multilinguality. *arXiv preprint arXiv:2406.10602*, 2024.

Ethan He, Abhinav Khattar, Ryan Prenger, Vijay Korthikanti, Zijie Yan, Tong Liu, Shiqing Fan, Ashwath Aithal, Mohammad Shoeybi, and Bryan Catanzaro. Upcycling large language models into mixture of experts. *arXiv preprint arXiv:2410.07524*, 2024.

Paul Heggarty, Warren Maguire, and April McMahon. Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1559):3829–3843, 2010.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Hukai Huang, Shenghui Lu, Yahui Shan, He Qu, Fengrun Zhang, Wenhao Guan, Qingyang Hong, and Lin Li. Dynamic language group-based MoE: Enhancing code-switching speech recognition with hierarchical routing. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, 2025.

J Jaavid, Raj Dabre, M Aswanth, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. RomanSetu: Efficiently unlocking multilingual capabilities of large language models via Romanization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15593–15615, 2024.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

João Maria Janeiro, Belen Alastruey, Francisco Massa, Maha Elbayad, Benjamin Piwowarski, Patrick Gallinari, and Loic Barrault. Mixture of languages: Improved multilingual encoders through language grouping. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 29695–29710, 2025.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Shareef Babu Kalluri, Prachi Singh, Pratik Roy Chowdhuri, Apoorva Kulkarni, Shikha Baghel, Pradyoth Hegde, Swapnil Sontakke, Deepak K T, S.R. Mahadeva Prasanna, Deepu Vijayasenan, and Sriram Ganapathy. The Second DISPLACE Challenge: DIarization of SPeaker and LAnguage in Conversational Environments. In *Interspeech 2024*, pp. 1630–1634, 2024. doi: 10.21437/Interspeech.2024-1833.

Siva Kalyan and Alexandre François. Freeing the comparative method from the tree model: A framework for historical glottometry. *Senri Ethnological Studies*, 98:59–89, 2018.

Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, et al. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages. *arXiv preprint arXiv:2511.09690*, 2025.

Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H Hoang, Eric Khiu, A Seza Doğruöz, and En-Shiun Annie Lee. URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the International Conference on Computational Linguistics*, pp. 6937–6952, 2025.

Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training mixture-of-experts from dense checkpoints. In *International Conference on Learning Representations*, 2023.

Yoohwan Kwon and Soo-Whan Chung. MoLE: Mixture of language experts for multilingual automatic speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, 2023.

Jaeseong Lee, Seung-won Hwang, Wonpyo Park, and Mingi Ji. Breaking ReLU barrier: Generalized MoEfication for dense pretrained models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024.

Sangmin Lee, Woojin Chung, and Hong-Goo Kang. LAMA-UT: Language agnostic multilingual ASR through orthography unification and language-specific transliteration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24393–24401, 2025.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021.

Chong Li, Yingzhuo Deng, Jiajun Zhang, and Chengqing Zong. Group then scale: Dynamic mixture-of-experts multilingual language model. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1730–1754, 2025a.

Jiahong Li, Yiwen Shao, Jianheng Zhuo, Chenda Li, Liliang Tang, Dong Yu, and Yanmin Qian. Efficient multilingual ASR finetuning via LoRA language experts. In *Interspeech 2025*, pp. 1138–1142, 2025b.

Xinjian Li, Florian Metze, David R Mortensen, Shinji Watanabe, and Alan W Black. Zero-shot learning for grapheme to phoneme conversion with language ensemble. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2106–2115, 2022.

Yunlong Liang, Fandong Meng, and Jie Zhou. THOR-MoE: Hierarchical task-guided and context-responsive routing for neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21433–21445, 2025.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 8–14, 2017.

Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Ayyoob ImaniGooghari, Orgest Xhelili, Haotian Ye, Chunlan Ma, François Yvon, and Hinrich Schütze. How transliterations improve crosslingual alignment. In *Proceedings of the International Conference on Computational Linguistics*, pp. 2417–2433, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Luis Lugo and Valentin Vielzeuf. Sustainable self-supervised learning for speech representations. *arXiv preprint arXiv:2406.07696*, 2024.

Guodong Ma, Wenxuan Wang, Yuke Li, Yuting Yang, Binbin Du, and Haoran Fu. LAE-ST-MOE: Boosted language-aware encoder using speech translation auxiliary task for E2E code-switching ASR. In *IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 1–8. IEEE, 2023.

Glenn W Milligan and Martha C Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

Bingshen Mu, Kun Wei, Qijie Shao, Yong Xu, and Lei Xie. HDMoLE: Mixture of LoRA experts with hierarchical routing and dynamic thresholds for fine-tuning LLM-based ASR models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, 2025.

Thomas Niesler et al. A first south african corpus of multilingual code-switched soap opera speech. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. Attentive statistics pooling for deep speaker embedding. In *Interspeech 2018*, pp. 2252–2256, 2018.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 914–921. IEEE, 2021.

Ankita Pasad, Bowen Shi, and Karen Livescu. Comparative layer-wise analysis of self-supervised speech models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023.

Juan N Pava, Caroline Meinhardt, Haifa Badi Uz Zaman, Toni Friedman, Sang T Truong, Daniel Zhang, Vukosi Marivate, and Sanmi Koyejo. Mind the (language) gap: Mapping the challenges of LLM development in low-resource language contexts. *Stanford Institute for Human-Centered Artificial Intelligence (HAI): Stanford, CA, USA*, 2025.

Yifan Peng, Muhammad Shakeel, Yui Sudo, William Chen, Jinchuan Tian, Chyi-Jiunn Lin, and Shinji Watanabe. OWSM v4: Improving open Whisper-style speech models via data scaling and cleaning. In *Interspeech 2025*, pp. 2225–2229, 2025.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 487–503, 2021.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.

Alan Saji, Jaavid Aktar Husain, Thanmay Jayakumar, Raj Dabre, Anoop Kunchukuttan, and Ratish Puduppully. RomanLens: The role of latent Romanization in multilinguality in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26410–26429, 2025.

Jiatong Shi, Dan Berrebbi, William Chen, En-Pei Hu, Wei-Ping Huang, Ho-Lam Chung, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Shinji Watanabe. ML-SUPERB: Multilingual Speech Universal PERformance Benchmark. In *Interspeech 2023*, pp. 884–888, 2023. doi: 10.21437/Interspeech.2023-1316.

Jiatong Shi, Shih-Heng Wang, William Chen, Martijn Bartelds, Vanya Bannihatti Kumar, Jinchuan Tian, Xuankai Chang, Dan Jurafsky, Karen Livescu, Hung yi Lee, and Shinji Watanabe. ML-SUPERB 2.0: Benchmarking Multilingual Speech Models Across Modeling Constraints, Languages, and Datasets. In *Interspeech 2024*, pp. 1230–1234, 2024. doi: 10.21437/Interspeech.2024-2248.

Igor Szöke, Miroslav Skácel, Ladislav Mošner, Jakub Paliesek, and Jan Černockỳ. Building and evaluation of a real room impulse response dataset. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):863–876, 2019.

Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics*, volume 19, pp. 035081. Acoustical Society of America, 2013.

Wenxuan Wang, Guodong Ma, Yuke Li, and Binbin Du. Language-routing mixture of experts for multilingual and code-switching speech recognition. In *Interspeech 2023*, pp. 1389–1393, 2023.

Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 4438–4450, 2020.

Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

Xun Wu, Shaohan Huang, and Furu Wei. Mixture of LoRA experts. In *International Conference on Learning Representations*, 2024.

Yifan Yang, Jianheng Zhuo, Zengrui Jin, Ziyang Ma, Xiaoyu Yang, Zengwei Yao, Liyong Guo, Wei Kang, Fangjun Kuang, Long Lin, et al. k2SSL: A faster and better framework for self-supervised speech representation learning. In *IEEE International Conference on Multimedia and Expo*, pp. 1–6. IEEE, 2025.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Are Transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. mHuBERT-147: A compact multilingual HuBERT model. In *Interspeech 2024*, pp. 3939–3943, 2024.

Shucong Zhang, Titouan Parcollet, Rogier van Dalen, and Sourav Bhattacharya. Linear-complexity self-supervised learning for speech processing. In *Interspeech 2024*, pp. 3480–3484, 2024.

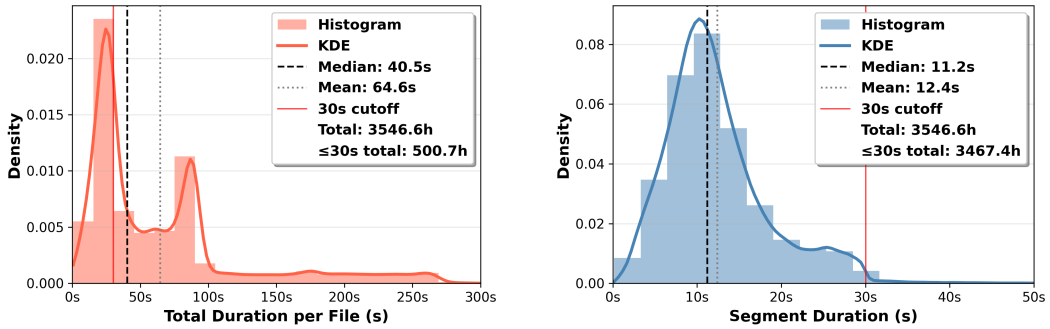
Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. Google USM: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*, 2023.

Guorui Zheng, Xidong Wang, Juhao Liang, Nuo Chen, Yuping Zheng, and Benyou Wang. Efficiently democratizing medical LLMs for 50 languages via a mixture of language family experts. *arXiv preprint arXiv:2410.10626*, 2024.

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Weihang You, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, et al. Opportunities and challenges of large language models for low-resource languages in humanities research. *arXiv preprint arXiv:2412.04497*, 2024.

Jian Zhu, Cong Zhang, and David Jurgens. Byt5 model for massively multilingual grapheme-to-phoneme conversion. 2022. doi: 10.48550/ARXIV.2204.03067. URL https://arxiv.org/abs/2204.03067.

Jian Zhu, Farhan Samir, Eleanor Chodroff, and David R Mortensen. Zipa: A family of efficient models for multilingual phone recognition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19568–19585, 2025.



(a) Duration of the original dataset.

(b) Duration of the dataset after segmentation.

Figure 5: Duration distributions in the omnilingual ASR corpus, aggregated across all splits.

A Dataset Details and Preprocessing Strategy

A.1 Dataset Details

FLEURS. FLEURS is a multilingual dataset encompassing 102 languages, specifically curated for ASR and speech translation. The corpus consists of read speech from Wikipedia articles, providing relatively clean recordings. Each language contains approximately 10 hours of supervised data, offering a balanced foundation for high-quality acoustic modeling.

CommonVoice. CommonVoice is a large-scale crowdsourced dataset covering 131 languages, exhibiting substantial acoustic variability, including diverse microphone qualities and background noise levels. We use the official splits from CommonVoice 22.0.

Omnilingual ASR Corpus. Omnilingual ASR corpus is a massively multilingual dataset spanning 348 languages, with a significant emphasis on low-resource languages and long-form speech. To mitigate training instability associated with long utterances, we perform audio segmentation via forced alignment.

A.2 Preprocessing Details of Omnilingual ASR Corpus

Dataset Characteristics. The Omnilingual ASR corpus differs from conventional ASR datasets in that it captures naturally occurring speech by prompting speakers with over 1,500 queries and recording their responses. As a result, it contains many long utterances exceeding 30 seconds, as shown in Figure 5. Rather than discarding them, which would significantly reduce the data, we apply a segmentation strategy to preserve the data capacity.

Segmentation Strategy. To address the aforementioned issue, we adopt a three-stage segmentation pipeline for utterances longer than 30 seconds: (1) segmentation using hard punctuation (e.g., periods, question marks, exclamation marks), (2) segmentation using soft punctuation (e.g., commas), and (3) forced segmentation based on a character limit. While most utterances can be segmented within 30 seconds using punctuation, we observe that some languages—particularly African languages—often contain long, punctuation-free running sentences. To address this, we apply forced segmentation at approximately 150 characters, splitting at the nearest word boundary. All segmentation is performed using MMS-FA. After segmentation, we filter samples based on alignment scores on a per-language basis, pruning low-quality segments to ensure overall data quality.

B Utilization of Romanization as a Phonetic Proxy

B.1 Representation Analysis in Multilingual ASR

Recent multilingual speech benchmarks, such as ML-SUPERB (Shi et al., 2023; 2024), have demonstrated the effectiveness of self-supervised speech models for multilingual ASR

across diverse languages. Most existing studies, however, primarily focus on improving downstream recognition performance through larger pretrained models, scaling strategies, or multilingual evaluation benchmarks.

In contrast, relatively little attention has been paid to how different transcription representations influence multilingual speech modeling. In particular, it remains unclear whether language-aware modeling techniques exhibit different behaviors when operating on orthographic or phonetic representations. Since our goal is not to introduce a new ASR backbone but to understand the effectiveness of language-aware expert grouping, we intentionally analyze MoLGE under two complementary representation spaces.

B.2 Romanization as a Scalable Alternative

Ideally, phonetic representations would be constructed using language-specific phoneme sequences, such as IPA-based transcriptions. In this context, we also considered adopting multilingual phonetic ASR benchmarks, such as ZIPA (Zhu et al., 2025) and PRiSM (Bharadwaj et al., 2026). However, constructing reliable phonetic transcriptions at the scale considered in this work (495 languages) remains impractical.

The primary challenge lies in multilingual grapheme-to-phoneme (G2P) conversion. Existing multilingual G2P toolkits provide only limited language coverage. For example, Phonemizer (Bernard & Titeux, 2021) and CharsiuG2P (Zhu et al., 2022) support approximately 120 and 100 languages, respectively, while broader-coverage systems such as Transphone (Li et al., 2022) rely extensively on zero-shot G2P generation, making phonetic quality difficult to guarantee for many low-resource languages. Consequently, generating consistent, high-quality phoneme transcriptions across hundreds of languages is currently infeasible.

Romanization provides a practical alternative that can be consistently applied across a substantially broader range of languages while preserving meaningful phonetic information. Consequently, it has been adopted in several large-scale multilingual speech systems, including MMS, where Romanized transcriptions are used for nearly one thousand languages. Although Romanization is not equivalent to true phonetic representations, prior studies have demonstrated that it serves as an effective phonetic proxy for multilingual speech modeling.

C Code-Switching and MoLGE

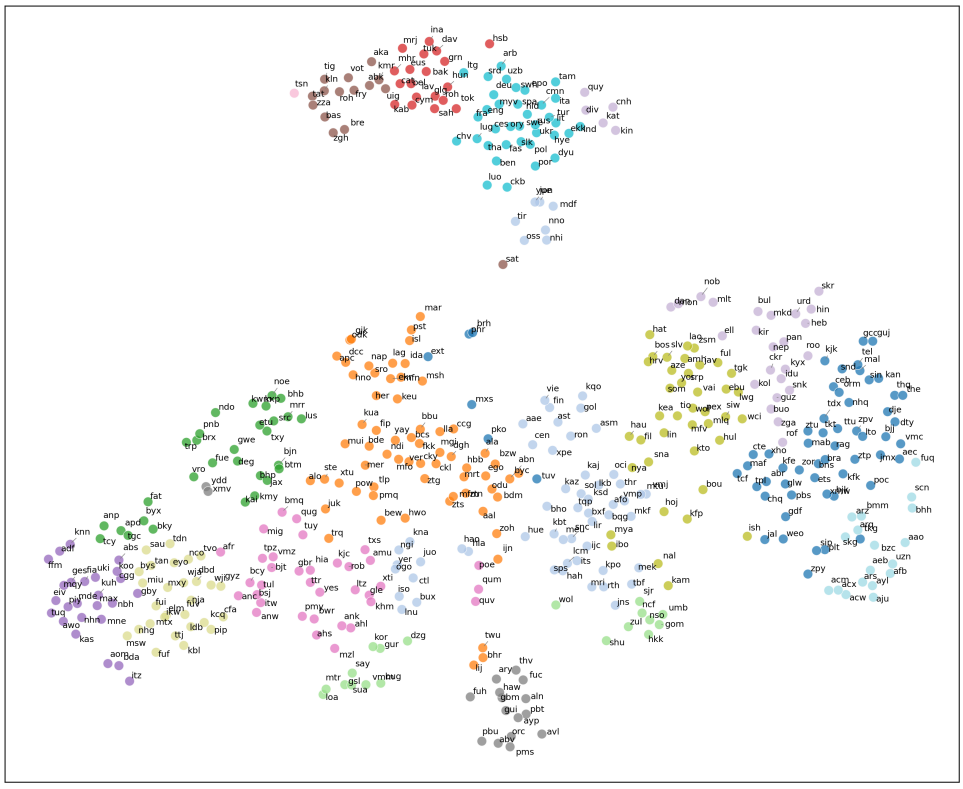
To further investigate the linguistic behavior of MoLGE under code-switching scenarios, we analyze two public corpora with time-aligned language annotations: DISPLACE 2024 (Kalluri et al., 2024) and the South African Soap Opera corpus (Niesler et al., 2018) The analysis indicates that, under sentence-level routing, the assigned expert group is largely determined by the matrix language, with agreement rates of 78.29% and 69.91% on DISPLACE 2024 and the South African Soap Opera corpus, respectively. In contrast, frame-level routing does not exhibit a clear or consistent language-dependent grouping pattern.

This behavior is likely attributable to the receptive field of the S3M backbone, where each frame represents approximately 25ms of acoustic context. Such short temporal windows are primarily optimized for capturing local phonetic information, making it difficult to reliably infer language identity at the frame level, particularly in code-switching speech.

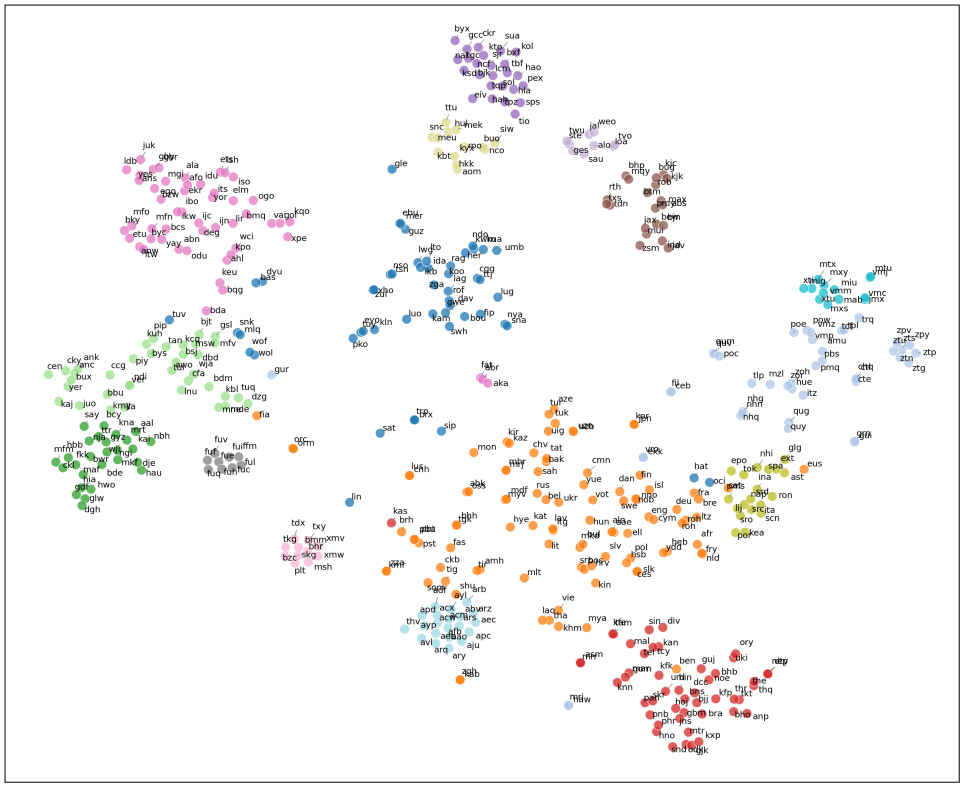
D Detailed Visualization on Language Clusters

In this section, we provide detailed visualizations of the clustering results for each explicit grouping strategy, as illustrated in Figures 6 and 7. Languages assigned to the same group are indicated by the same color. Each plot is annotated with ISO 639-3 language codes following the Ethnologue[4] convention.

[4]https://www.ethnologue.com/

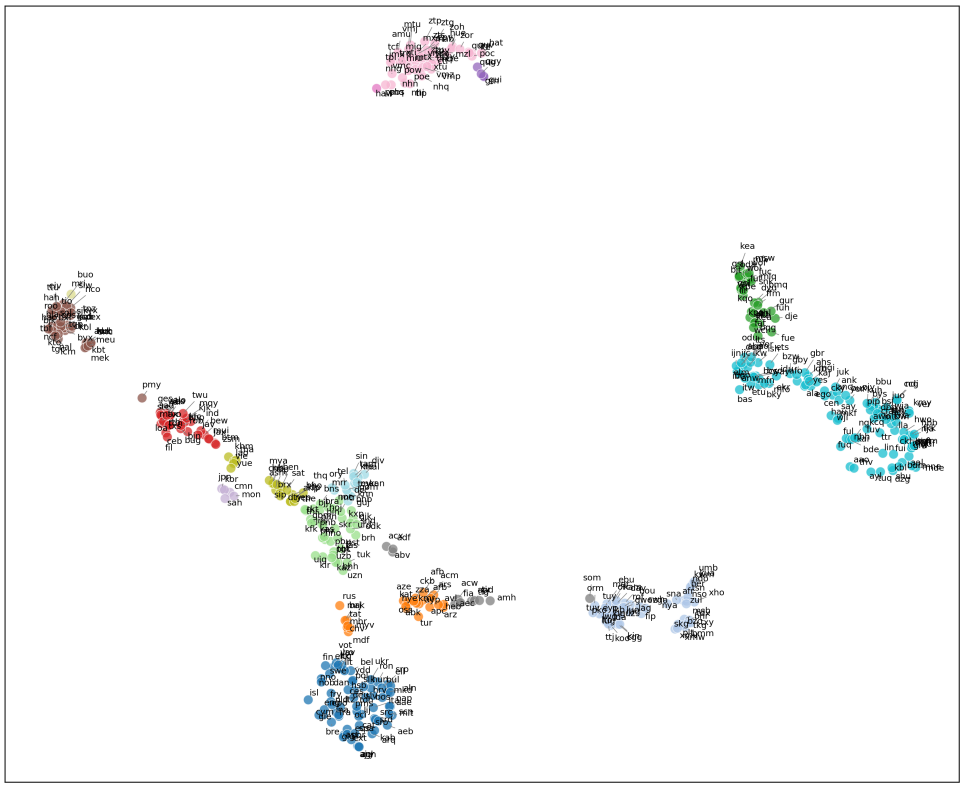


(a) Cluster formation using the SSL embedding.

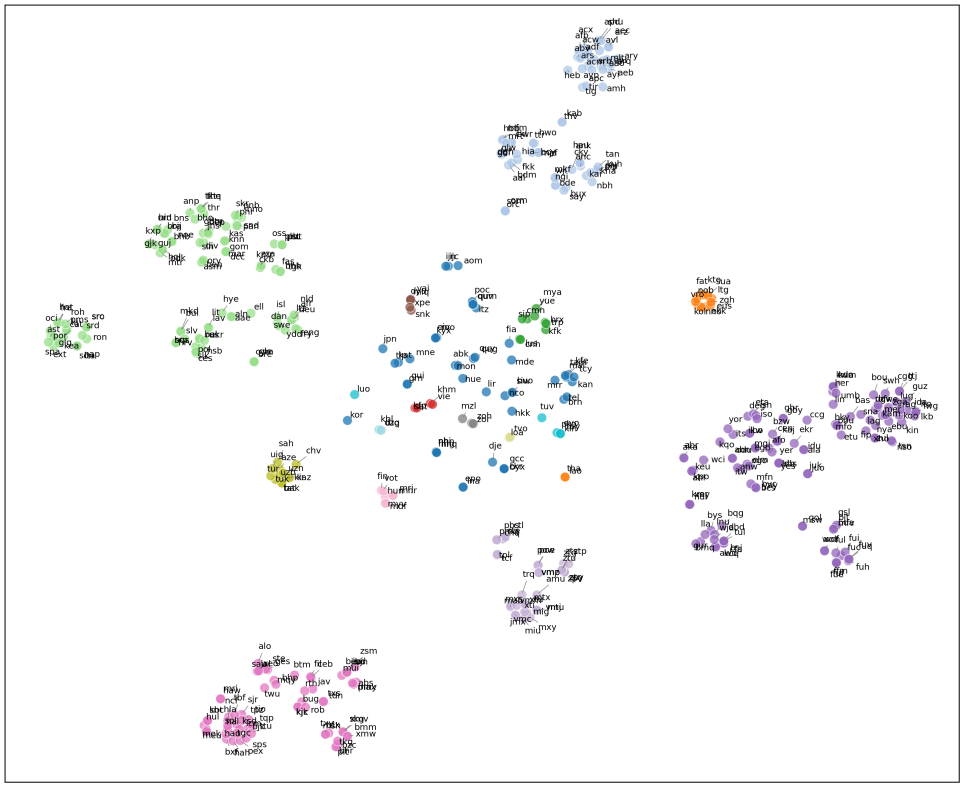


(b) Cluster formation using the LID embedding.

Figure 6: Comparison of embedding-guided clusters.



(a) Cluster formation using the geographic vector.



(b) Cluster formation using the genetic vector.

Figure 7: Comparison of knowledge-guided clusters.