PRiSM: Benchmarking Phone Realization in Speech Models

Anonymous ACL submission

Abstract

Phone recognition (PR) serves as the atomic interface for language-agnostic modeling for cross-lingual speech processing and phonetic analysis. Despite prolonged efforts in developing PR systems, current evaluations only measure surface-level transcription accuracy. We introduce PRiSM, the first open-source benchmark designed to expose blind spots in phonetic perception through intrinsic and extrinsic evaluation of PR systems. PRiSM standardizes transcription-based evaluation and assesses downstream utility in clinical, educational, and multilingual settings with transcription and representation probes. We find that diverse language exposure during training is key to PR performance, encoder-CTC models are the most stable, and specialized PR systems still outperform LALMs. PRiSM releases code, recipes, and datasets to move the field toward multilingual speech models with robust phonetic ability.

1 Introduction

Phone recognition (PR) entails transcribing speech into phonetic units that capture the physical realization of sounds, independent of language-specific phonological constraints. By preserving acoustic nuances often abstracted away by word- or phoneme-level models[1], PR provides a robust foundation for cross-lingual speech processing (Li et al., 2022; Yusuyin et al., 2025) and downstream applications in clinical (Shriberg et al., 2025; Choi et al., 2025) and educational settings (Tu et al., 2018; Inceoglu et al., 2023).

PR models have scaled substantially to cover diverse linguistic settings (see § 2.1), yet existing evaluations remain difficult to compare across studies. For example, models often differ in language coverage and phone inventories (Zhu et al., 2025), and evaluation metrics are not standardized (Li et al., 2025). A common response has been to fix a metric (Taguchi et al., 2023; Li et al., 2025) and expand the number of test datasets to mitigate bias (Zhu et al., 2025). Yet this approach scales poorly due to the scarcity of phonetically transcribed data. Moreover, transcription error rates do not necessarily reflect a model's phonetic capabilities or practical utility. Error rates in PR are inherently noisier than in ASR, as phones, unlike lexical units, correspond to a lower-level, articulatorily defined abstraction of the acoustic signal.

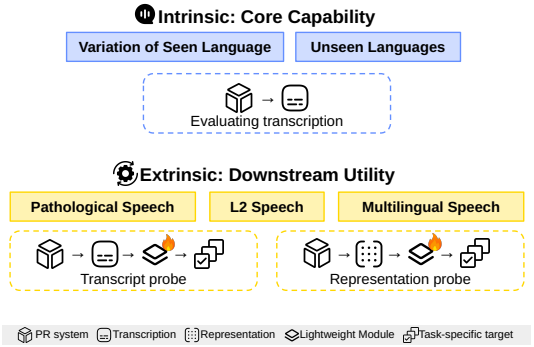


Figure 1: PRiSM is the first open-source benchmark for phone recognition systems, covering intrinsic and extrinsic evaluations, i.e., transcription task and downstream task performance.

Furthermore, the link between transcription accuracy and downstream performance is often assumed rather than empirically proven. In practice, models leverage phonetic information via two channels: explicit transcriptions and latent internal representations. The latter are especially potent, as they encode rich phonetic cues (see § 2.2). Consequently, metrics based solely on transcription error fail to capture the full utility and nuanced quality of these representations.

Therefore, we propose **PRiSM** to fairly benchmark **P**hone **R**ealization **i**n **S**peech **M**odels. PRiSM assesses PR systems[2] intrinsically through

[1] For example, *tell* may be transcribed as [tʰɛɬ] in Mainstream American English and [tʰɛl] in Scottish English, while the phonemic form of *tell* is consistently /tɛl/.

[2] Any pipeline that converts speech into phonetic units.

transcription error, and extrinsically through utility in clinical, educational, and multilingual speech tasks using generated transcriptions and hidden representations. PRiSM applies to PR systems ranging from specialized PR models to general speech-to-text (S2T) systems, including Large Audio Language Models (LALMs), which are increasingly used for general speech tasks despite limited evaluation of their phonetic abilities (Peng et al., 2024; Arora et al., 2025).

PRiSM is the first open-source benchmark for PR systems, for which code, evaluation recipes, and datasets are released, where licensing permits. With its reproducible and expandable framework, PRiSM supports researchers in understanding model behavior and training strategies, and helps practitioners make informed model choices. We evaluate a broad range of PR systems and find that: (i) **language exposure matters**: Seen languages benefit from familiar patterns, unseen from multilingual training; (ii) **data and architecture shape performance**: Broad, diverse coverage improves results, while encoder-CTC architectures are more stable; (iii) **LALMs lag behind specialized PR models**.

Ultimately, our goal is to establish a common evaluation basis to drive progress toward PR systems that capture robust and generalizable phonetic information across resource conditions.

2 Background

2.1 Phone Recognition Systems

PR can be viewed as a variant of the S2T task that maps speech to phonetic symbols such as IPA (International Phonetic Association, 1999). In this work, we use "PR system" to refer broadly to any system capable of converting speech into IPA in a language-agnostic fashion.

Modern PR systems are typically fine-tuned from ASR systems (Baevski et al., 2020; Radford et al., 2023) or trained from scratch on ASR datasets (Zhu et al., 2024) with transcriptions automatically converted to IPA using grapheme-to-phoneme (G2P) tools (Mortensen et al., 2018; Zhu et al., 2022). Language-specific approaches (Li et al., 2020; Gao et al., 2021) rely on phoneme inventories, while language-agnostic approaches, which we focus on, seek to learn phonetic representations generalized across languages. LALMs have recently become prominent in speech tasks and have shown competitive performance with cascaded systems that combine LLMs with speech processing modules (Yang et al., 2025), motivating interest in their application to PR (Huang et al., 2025; Wang et al., 2025). We describe the systems investigated in this work in § 4.

2.2 Phonetic information in PR systems

Explicitly generated phonetic transcriptions are easy for humans to inspect and utilize. For example, faithful phonetic transcriptions of the speech of a child with a speech sound disorder can help a clinician understand the nature of the disorder and design interventions (Dodd, 2013). Nevertheless, representing continuous speech with discrete symbols inherently incurs information loss, filtering out non-linguistic variation.

Internal model representations serve as a complement that retains richer information. Speech models trained on S2T tasks produce temporally aligned representations that capture empirically useful acoustic-phonetic (Choi et al., 2024), articulatory (Cho et al., 2024), and even semantic (Ma et al., 2025) features. The most widely used S2T model representations are from end-to-end ASR models such as Whisper (Radford et al., 2023) and WavLM (Chen et al., 2022). In contrast, LALMs' representations are often inaccessible or difficult to analyze, as they focus mainly on textual output and lack strict temporal alignment with input speech.

2.3 Assessing phonetic/phonological ability

In the text modality, language models are evaluated with text input and output. PhonologyBench (Suvarna et al., 2024) evaluates G2P, syllable counting, and rhyme judgment, while Bunzeck et al. (2025) and Goriely and Buttery (2025) probe phonological knowledge using minimal pairs and word segmentation. In the speech modality, models are evaluated with speech (and optionally text) input or representation output. SUPERB (Yang et al., 2021) and Dynamic-SUPERB (Huang et al., 2025) include phoneme recognition, phonological feature analysis, and pronunciation evaluation. BabySLM (Lavechin et al., 2023) and the ZeroSpeech challenges (Nguyen et al., 2020) propose metrics that evaluate phonological and acoustic-phonetic contrasts based on minimal pairs. In contrast to previous work, PRiSM evaluates phonetic ability in both text and speech through intrinsic and extrinsic tasks.

Abbr.	Task	Dataset	Lang.
Intrinsic: Core Capability (Metrics ↓ Lower is better)			
Phone Recognition (PFER)			
PR-tmt	Variation of Seen Language	TIMIT (Garofolo et al., 1993)	English
PR-arc	Variation of Seen Language	L2-ARCTIC Perceived (Zhao et al., 2018b)	English
PR-saa	Variation of Seen Language	Speech Accent Archive (Weinberger, 2015)	English
PR-drc	Unseen Languages	DoReCo (Paschen et al., 2020)	45 langs
PR-vox	Unseen Languages	VoxAngeles (Chodroff et al., 2024)	95 langs
PR-tsm	Unseen Languages	Tusom2021 (Mortensen et al., 2021)	Tusom
Extrinsic: Downstream Utility (Metrics ↑ Higher is better)			
Pathological Speech: Dysarthria Intelligibility Prediction (τ) & Child Speech Disorder Detection (F1)			
DYS-ez	Dysarthria Intelligibility Prediction	EasyCall (Turrisi et al., 2021)	Italian
DYS-ua	Dysarthria Intelligibility Prediction	UASpeech (Kim et al., 2008)	English
CSD-us	Child Speech Disorder Detection	UltraSuite (Eshky et al., 2018)	English
L2 Speech: L1 Classification (F1) & L2 Assessment (τ)			
L1-eda	L1 Classification	EdAcc (Sanabria et al., 2023)	English
L1-arc	L1 Classification	Kominek and Black (2004) & Zhao et al. (2018b)	English
L2-so	L2 Assessment	Speechocean762 (Zhang et al., 2021a)	English
Multilingual: Lang. ID (F1), Geolocation (Recall@1) & Phone Inventory Induction (F1-PI)			
LID-fl	Lang. ID (LID)	FLEURS-24 (Conneau et al., 2023)	24 langs
GEO-v	Speech Geolocation	Vaani (Ghosh et al., 2025)	Hindi Dialects
PI-drc	Phone Inventory Induction	DoReCo (Paschen et al., 2020)	45 langs

Table 1: List of evaluation tasks. Blue denotes core capabilities, where lower scores are better. Yellow denotes downstream utility, where higher scores are better. *F1-PI* is described in § B.1. See Appendix A for license details.

3 Evaluation Framework of PRiSM

PRiSM covers intrinsic (§ 3.1) and extrinsic (§ 3.2) evaluations shown in Figure 1. Intrinsic evaluation compares predicted transcriptions to gold labels, while extrinsic evaluation measures transcriptions and internal representations on downstream tasks. In extrinsic evaluation, transcriptions provide a direct and interpretable signal of explicit phonetic content, whereas representations are commonly used in downstream tasks but may encode non-phonetic information. Table 1 summarizes included datasets and metrics.

3.1 Intrinsic: Core Capability

We use **Phonetic Feature Error Rate (PFER)** to measure the distance between reference and predicted transcriptions. Unlike Phone Error Rate (PER), which treats each phone as a token, PFER computes the edit distance $D(\cdot,\cdot)$ over articulatory features $\text{feat}(\cdot)$ such as roundness or voicing. As shown in Equation 1, where u denotes an utterance (sequence of phones where i indexes this sequence) and u^* its ground truth, PFER is calculated as the total feature edit distance across all utterances divided by the total number of phones, representing the percentage of incorrect features (Mortensen et al., 2016).

$$\text{PFER} = \frac{1}{\sum_i |u_i^*|} \sum_i D(\text{feat}(u_i^*), \text{feat}(u_i)) \quad (1)$$

The tasks comprise two categories: **Variation of seen languages** includes regional and non-native speech, testing whether PR systems rely excessively on seen patterns rather than the actual input. **Unseen languages** assess the system's language-agnostic phonetic knowledge. Details of each task and dataset are in § A.1.

3.2 Extrinsic: Downstream Utility

We evaluate PR systems using two downstream probes, namely the transcript probe and the representation probe. For **transcript probe (TP)**, input consists of predicted phonetic transcriptions, and the probe is a text-based bi-GRU. For **representation probe (RP)**, following the setup in Turian et al. (2022), we use the last layer's hidden representations as input and temporal pooling with attention followed by a Multi-Layer Perceptron as the probe. Metrics for each task are listed in Table 1 and the detailed experimental setup is in Appendix C.

We consider three categories of downstream tasks where phonetic information is essential. In

Model	Architecture (Enc / Dec)	Loss	SSL Pre-training Data	Phone Recognition Data	Langs
W2V2P-LV60 (Xu et al., 2022)	Enc: Wav2Vec2	CTC	LibriLight	MLS, CV, Babel	40+
W2V2P-XLSR53 (Xu et al., 2022)	Enc: XLSR53	CTC	MLS, CV, Babel	MLS, CV, Babel	40+
MultiIPA (Taguchi et al., 2023)	Enc: XLSR53	CTC	MLS, CV, Babel	CV 11.0	7
ZIPA-CTC (Zhu et al., 2025)	Enc: Zipformer	CR-CTC	None	IPAPack++	88
ZIPA-CTC-NS (Zhu et al., 2025)	Enc: Zipformer	CR-CTC	None	IPAPack++ & PL	4k
POWSM (Li et al., 2025)	Enc: E-Branchformer Dec: Transformer	CTC-Att	None	IPAPack++	88
POWSM-CTC (ours)	Enc: E-Branchformer	Int-CTC	None	IPAPack++	88
Gemini 2.5 Flash (Comanici et al., 2025)	Closed	N/A	Closed	Closed	>200
Qwen3-Omni-Instruct (Xu et al., 2025)	Enc: AuT Dec: MoE Transformer	AR	Closed	Closed	19

Table 2: Included PR systems. Architecture abbrv.: Encoder (Enc), Decoder (Dec), Audio Transformer (AuT), Mixture-of-Experts (MoE); Loss abbrv.: Consistency Regularized CTC (CR-CTC) (Yao et al., 2025), Hybrid CTC/Attention (CTC-Att) (Watanabe et al., 2017), Intermediate CTC (Int-CTC) (Lee and Watanabe, 2021), Autoregressive (AR); Data abbrv.: Multilingual LibriSpeech (MLS), Common Voice (CV), Pseudo-labeled (PL).

pathological speech assessment, phonetic transcriptions are used to document patients' speech and support diagnosis and treatment planning (Ball et al., 2009; Nelson et al., 2020). In **L2 speech assessment**, phonetic cues enable pronunciation feedback (Franco et al., 2010) and accent classification (Angkititrakul and Hansen, 2006). In **multilingual speech identification**, analyzing phonetic and phonological differences across languages and dialects, such as phone inventories, phonotactics, and phoneme realization, is crucial (Schultz and Kirchhoff, 2006). We describe each task and dataset in detail in § A.2.

4 Benchmarked Models

Table 2 summarizes the studied model families:

- **Wav2Vec2Phs**: MultiIPA, W2V2P-LV60, and W2V2P-XLSR53 are fine-tuned variants of Wav2Vec2 (Baevski et al., 2020), contrastively pre-trained speech SSL models, and differ in pre-training coverage and phone recognition fine-tuning datasets.
- **ZIPAs**: ZIPA-CTC and ZIPA-CTC-NS are encoder-CTC models trained from scratch on multilingual data, with ZIPA-CTC-NS further trained on large-scale pseudo-labeled data from ZIPA-CTC.
- **POWSMs**: POWSM is an attention-based encoder-decoder (AED) model trained on the same dataset as the ZIPAs and augmented for other S2T tasks. Following their framework, we also train POWSM-CTC, an encoder-CTC variant for comparison.
- **LALMs**: We include Gemini 2.5 Flash (closed-source) and Qwen3-Omni-Instruct (open-weight), both state-of-the-art systems widely used in recent studies (Lee et al., 2025). Since their representations are not easy to access or pool, we probe them with zero-shot prompting, which is a form of context-based fine-tuning (Petrov et al., 2023). The prompts are in Appendix D.
- **Other baselines**: We include a naive baseline that randomly predicting the class or the most frequent location (`GEO-v`). We also include WavLM[3](Chen et al., 2022) and Whisper[4](Radford et al., 2023) as competitive baselines for representation probing.

5 Results and Discussion

Table 3 presents the evaluation of PR. Table 4 presents a comprehensive breakdown of downstream evaluations. In general, ZIPA-CTC-NS performs well in all settings, while Whisper excels in RP. LALMs generally remain less competitive.

5.1 Intrinsic Evaluation

We observe a consistent trend for language variation: CTC-based models generally outperform LALMs, followed by AED models. For MultiIPA, English appears during pretraining but not fine-tuning, highlighting the importance of language coverage in PR data. POWSM performs poorly on `PR-saa`, likely due to decoder search issues on long speech sequences; meanwhile, a text-based G2P model (Zhu et al., 2022) achieves a PFER of 10.2, beating Gemini 2.5 Flash despite modeling only canonical pronunciations.

[3] `https://huggingface.co/microsoft/wavlm-base`
[4] `https://huggingface.co/openai/whisper-small`

Model	Variation of Seen Language				Unseen Languages			
	PR-tmt	PR-arc	PR-saa	Avg.	PR-drc	PR-vox	PR-tsm	Avg.
MultiIPA*	16.3	15.5	13.8	15.2	18.3	15.2	30.5	21.3
W2V2P-LV60	13.2	10.9	9.4	11.2	17.8	15.7	24.9	19.5
W2V2P-XLSR53	13.5	9.9	9.0	10.8	17.3	**13.9**	31.9	21.0
ZIPA-CTC	**13.1**	**9.7**	9.0	**10.6**	18.0	17.0	23.7	19.6
ZIPA-CTC-NS	**13.1**	**9.7**	**8.9**	**10.6**	**16.8**	17.1	23.1	19.0
POWSM	13.7	11.3	27.6	17.5	17.1	17.1	**22.0**	**18.7**
POWSM-CTC	**13.1**	10.3	10.0	11.1	18.1	15.3	32.2	21.9
Gemini 2.5 Flash**	15.2	12.7	13.2	13.7	105.3	19.7	36.3	53.8
Qwen3-Omni-Instruct**	15.1	11.9	9.1	12.0	150.2	49.0	117.1	105.4

Table 3: PFER of the intrinsic evaluation (↓). *English is included during pretraining but not fine-tuning. **Some of the "unseen languages" may have appeared in the training data. See § 5.1 for details.

Model	Pathological Speech			L2 Speech			Multilingual Speech			Score
	DYS-ez	DYS-ua	CSD-us	L1-eda	L1-arc	L2-so	LID-fl	GEO-v	PI-drc	
Naive Baseline	0.7 ±1.6	-0.8 ±0.9	41.8 ±1.0	6.3 ±0.4	14.3 ±0.3	1.5 ±1.0	4.3 ±0.3	3.3 ±0.0	—	8.9
Transcript Probe (TP)										
MultiIPA	48.2 ±0.2	45.6 ±1.4	93.6 ±1.6	**10.0 ±1.0**	**50.5 ±0.9**	33.3 ±1.7	89.3 ±0.5	44.5 ±0.4	40.9	**44.3**
W2V2P-LV60	42.4 ±1.3	50.3 ±0.9	95.6 ±1.4	7.6 ±0.5	38.0 ±0.3	36.1 ±1.7	91.4 ±0.2	**45.7 ±0.9**	51.3	42.0
W2V2P-XLSR53	49.2 ±0.8	47.6 ±0.8	92.3 ±2.6	9.1 ±0.6	43.1 ±0.6	37.5 ±0.8	94.1 ±0.2	44.5 ±1.2	56.9	43.8
ZIPA-CTC	55.0 ±0.6	**57.0 ±0.5**	91.7 ±2.3	6.6 ±0.4	30.5 ±0.5	36.6 ±2.8	95.6 ±0.2	44.1 ±1.0	55.2	43.5
ZIPA-CTC-NS	**56.6 ±0.8**	51.1 ±1.3	**99.4 ±0.5**	6.7 ±0.3	30.0 ±0.3	**40.8 ±0.8**	**95.9 ±0.1**	44.7 ±1.8	56.6	44.2
POWSM	52.7 ±1.7	46.1 ±0.8	94.3 ±1.3	6.5 ±0.8	28.0 ±0.3	28.4 ±2.2	95.1 ±0.5	43.7 ±1.4	48.7	39.6
POWSM-CTC	53.3 ±0.4	46.5 ±0.6	96.9 ±0.7	6.4 ±0.5	29.8 ±0.1	26.8 ±0.7	90.4 ±0.4	42.9 ±0.8	**57.7**	40.2
Gemini 2.5 Flash	27.9 ±0.6	38.5 ±0.4	95.0 ±1.6	6.4 ±0.4	22.3 ±0.4	20.1 ±1.1	91.8 ±0.3	33.2 ±1.0	39.1	31.6
Qwen3-Omni-Instruct	52.5 ±1.8	49.4 ±1.0	98.9 ±0.8	6.9 ±0.6	30.5 ±0.3	15.6 ±1.4	89.3 ±0.2	34.7 ±1.2	44.5	39.2
Representation Probe (RP)										
MultiIPA	65.5 ±4.0	77.0 ±1.4	98.5 ±0.8	11.7 ±1.1	53.0 ±3.3	46.3 ±1.9	78.2 ±1.0	24.5 ±3.7	—	56.5
W2V2P-LV60	67.2 ±1.8	79.9 ±0.9	98.6 ±0.6	12.0 ±0.3	60.7 ±2.9	49.9 ±1.0	76.6 ±1.3	24.6 ±2.1	—	59.4
W2V2P-XLSR53	70.8 ±2.2	**82.0 ±2.2**	99.2 ±0.7	13.0 ±1.1	47.0 ±6.0	50.7 ±3.1	81.0 ±2.3	21.5 ±3.1	—	58.2
ZIPA-CTC	73.2 ±2.2	74.7 ±1.2	**99.5 ±0.3**	13.9 ±0.9	73.4 ±2.5	54.0 ±0.8	96.1 ±0.7	23.0 ±1.2	—	62.9
ZIPA-CTC-NS	71.2 ±2.2	75.1 ±0.8	98.6 ±0.9	13.7 ±0.6	74.1 ±2.8	54.3 ±0.5	**96.8 ±0.3**	24.0 ±1.5	—	62.7
POWSM	73.0 ±3.0	70.8 ±1.1	**99.5 ±0.3**	10.3 ±1.2	68.0 ±1.9	53.1 ±0.3	96.5 ±0.1	21.5 ±2.2	—	60.4
POWSM-CTC	73.6 ±1.5	66.7 ±1.6	97.9 ±0.9	8.0 ±0.7	53.0 ±0.5	45.7 ±3.0	75.4 ±1.5	14.1 ±2.6	—	55.2
WavLM	69.2 ±2.0	77.5 ±1.4	99.0 ±0.5	14.4 ±1.0	58.3 ±2.0	50.2 ±1.4	76.2 ±3.2	23.5 ±4.6	—	59.4
Whisper	**74.8 ±1.1**	79.5 ±0.3	**99.5 ±0.3**	**24.3 ±1.6**	**84.3 ±3.0**	**57.2 ±0.8**	96.3 ±0.5	**35.0 ±2.7**	—	**68.5**
Zero-shot										
Gemini 2.5 Flash	21.4	50.4	75.3	32.7	43.9	35.8	91.5	6.5	—	41.5
Qwen3-Omni-Instruct	27.0	61.7	70.9	18.2	31.8	49.8	59.1	5.3	—	41.5

Table 4: PR system performance on extrinsic tasks (↑). Results are reported as mean ± standard deviation across 5 random seeds where applicable. Best numbers are **bolded** and second-best underlined. See § 5.2 for details. The formula for aggregrated score is in § B.2.

For unseen languages, AED and CTC models show comparable performance, whereas LALMs perform poorly, sometimes producing repeated or degenerate outputs indicative of limited training exposure (Holtzman et al., 2020). Performance also varies across datasets, and language-wise breakdowns reveal heterogeneous behavior. POWSM outperforms POWSM-CTC and exhibits performance comparable to ZIPAs, suggesting that incorporating a degree of language modeling may improve generalization by capturing shared phonological patterns, as further analyzed in § 6.1.

These trends show that **variation in seen languages benefits from outputs grounded in known patterns**, whereas **unseen languages benefit from multilingual training and learned phonological patterns**.

5.2 Extrinsic Evaluations

For transcript probe, ZIPAs and W2V2P-XLSR53 are generally competitive. ZIPAs perform well on pathological speech, likely due to their normalized, smaller vocabularies, which approximate broad transcription known to be reliable for speech disorders (Shriberg and Lof, 1991), while W2V2P-XLSR53 benefits from diverse pretrain-

ing data. Multilingual training further improves performance, especially on multilingual tasks. We discuss `PI-drc` in § 6.2 as an example. For representation probe, Whisper is strong, suggesting that large-scale ASR pretraining produces representations that retain phonetic information.

A trade-off of TP and RP emerges among specialized PR models: for example, Wav2Vec2Phs achieve strong TP results on L2 speech but show limited gains on RP, whereas ZIPAs underperform on TP yet excel on RP. Task category also influences their relative performance: Pathological speech benefits more from RP, L2 speech falls in the middle, and multilingual tasks tend to favor TP. We hypothesize that transcripts act as a structured bottleneck: pathological speech relies on features such as timbre and prosody, whereas multilingual settings benefit less from acoustic detail. We investigate the behavior of TP on `GEO-v` in § 6.3.

LALMs show task-dependent performance. Notably, Qwen3-Omni-Instruct achieves competitive TP on pathological speech, but they generally perform poorly in zero-shot settings and underperform on languages other than English (`DYS-ez`). An exception is L2 speech, where the gap is smaller, explored in § 6.4.

Overall, our results highlight the **importance of evaluating PR systems with a combination of intrinsic and extrinsic tasks**. Intrinsic evaluation alone may not fully capture phonetic capabilities, while extrinsic evaluation reveals that relative performance on TP and RP is task-dependent. **Multilingual pretraining and fine-tuning improve performance** across model families, and **encoder-CTC based architectures provide more stable PR performance** in new domains. In contrast, **LALMs remain limited** in phone recognition and related tasks.

6 Analysis

We conduct several analyses to anchor our observations. In § 6.1, We examine how architectural choices affect the balance between phonotactics and acoustics, echoing with evaluation results. In § 6.2, we study multilingual generalization and confirm that encoder-only architectures trained with diverse language coverage at all stages perform well for PR. In § 6.3, we analyze TP in detail and show that it effectively captures phone distribution differences across regions. Finally, in § 6.4, we assess zero-shot performance of LALMs on challenging tasks, concluding that they remain insensitive to sociophonetic variation.

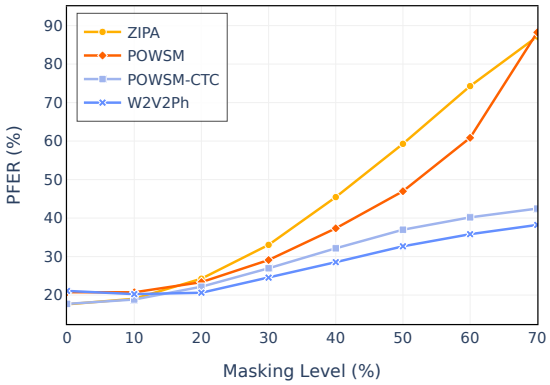


Figure 2: PFER vs Phone masking rate. A PR model that relies only on acoustics should produce a horizontal line. Encoder-only models trained with CTC loss retain acoustic fidelity at high masking levels. See § 6.1.

6.1 Phonotactics or the Acoustic Signal

Ideally, PR systems would faithfully transcribe the actual pronunciation in the speech signal via acoustic modeling. Instead, model transcriptions often normalize toward standard pronunciations or other probabilistically likely phone patterns (Zhu et al., 2025; Li et al., 2025), essentially relying on (phone-level) language modeling (Pimentel et al., 2020).[5] Additionally, models can also overfit phonotactics from the high-resource languages. In this experiment, we investigate the extent to which PR systems rely on such phonotactic patterns present in the training data, as opposed to information derived directly from acoustic signal.

Using TIMIT (Garofolo et al., 1993)'s time-aligned phone transcripts, we replace $p\%$ of phones with silence, transcribe the modified speech using PR, and compute PFER against a reference containing only the remaining phones. In Figure 2, we plot the phone masking rate against the PFER for different model families. For a model that only relies on the acoustic waveform for prediction, the curve would be a horizontal line. However, for models that rely on phonotactics, the PFER will increase with greater noise. While all models start at a similar PFER, Wav2Vec2Phs and POWSM-CTC perform better than ZIPAs and POWSM at higher masking levels. Thus **Wav2Vec2Phs relies more on the acoustic**

[5] A common example of such phonotactic knowledge is the intuition that *brick* [bɹɪk] is a valid phone sequence in English while *bnick* [bnɪk] is not (Chomsky and Halle, 1968).

signal than on phonotactics. While POWSM is an AED model trained with next-token prediction, Wav2Vec2Phs and POWSM-CTC are encoder-only models trained with the CTC objective.

However, ZIPAs are also encoder-only (Zipformer (Yao et al., 2024)) models, but they are trained with a consistency regularized CTC (CR-CTC) loss (Yao et al., 2025). The high PFER of ZIPA means that **CR-CTC loss biases the model to learn from both phonotactics and acoustics**. We hypothesize that ZIPAs' Zipformer-based downsampling bottleneck can be another possible reason for ZIPA's high PFER. We also observe that the insertion rates for different models follow the same trend as the curves in Figure 2, showing that POWSM and ZIPA produce phonetic transcriptions even when there is no input speech. Interestingly, ZIPA-CTC-NS and POWSM perform best on unseen languages, but POWSM struggles with seen-language variation, and ZIPA-CTC-NS underperforms on pathological speech. This aligns with the idea that some tasks benefit from learned phonological patterns, while others depend more on capturing acoustic information.

6.2 Zero-Shot Phonetic Inventory Induction

Identifying the inventory of phones in a new language is an important linguistic application and often an early step toward developing a standardized transcription system for it. Such a task requires PR models to recognize phones correctly in unseen phonetic environments. Therefore, it relies on the phonetic diversity the models have seen in the input speech signal during training. We explore these behaviors in this set of experiments.

Our dataset, derived from DoReCo, consists of low-resource languages absent from the training corpora of all models. The transcripts from all models are used to compute the phone inventory after applying PanPhon-based phone tokenization (Mortensen et al., 2016) followed by a set union over detected phones. The ground truth inventory is constructed similarly using the phonetic transcriptions provided by DoReCo and set similarity metrics (§ B.1) are computed. We show the macro-averaged values in Figure 3.

POWSM-CTC emerges as the strongest model. The large gap between POWSM-CTC and POWSM (which differ only in architecture) suggests that the **encoder-only architecture plays a crucial role in high precision transcripts even in an unseen phonetic environment**. As for ZIPAs, which differ in training data, ZIPA-CTC-NS is more precise than ZIPA-CTC. The extended multilingual training of ZIPA-CTC-NS on pseudo-labeled data leads to more precise phone predictions for unseen languages. This suggests that noisy pseudo-labels allow for improved precision for new languages. Similar trends is seen in comparing W2V2P-XLSR53 vs W2V2P-LV60 and MultiIPA, where multilingual SSL alone is insufficient for MultiIPA. **Essentially, broader language coverage in both pre-training and supervised training result in a more precise model.** Although the size of IPAPack++ (17k hr) is much smaller than that used for Wav2Vec2Phs (~160k hr), the larger number of languages in the supervised training stage (88 vs ~40) leads to better recall for ZIPAs, compared to Wav2Vec2Phs. This suggests that **diversity of languages is as important as the volume of data**. Most models have a high recall (> 70) and low precision (< 50), suggesting that most predicted phones are incorrect and predictions have a high entropy.

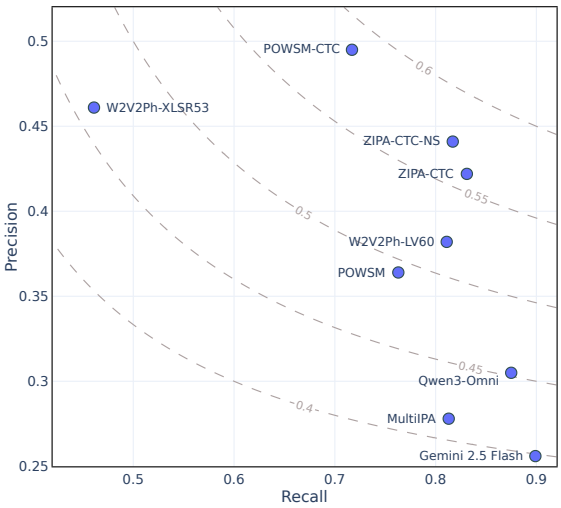


Figure 3: Precision and Recall scores of PR systems on phone inventory induction for unseen languages (§ 6.2). CTC models trained with highly multilingual data are more stable.

6.3 Geolocation for Dialectal Speech

In Table 4, we observe that our TPs significantly outperform the RPs on Hindi dialectal geolocation (Foley et al., 2024), where the former observes an average error of 146 km, while the latter observes an average error of 253 km. As a reference, our data is spread over 1478 km (East-West) and 1703 km (North-South), covering the entire Hindi speaking region of India. This per-

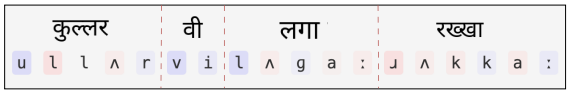


Figure 4: Attribution map from Vaani (Ghosh et al., 2025). Red supports and blue opposes correct geolocation. W2V2P-LV60 detects doubled phones (§ 6.3).

formance is surprising, as the cascade-based approach loses suprasegmental information such as intonation that provide strong phonetic cues for the differentiation of dialects (Vicenik and Sundara, 2013; Grabe and Post, 2002). However, our results provide empirical evidence that morphological and phonetic differences suffice for fine-grained differentiation between Hindi dialects (Gumperz, 1958).

We hypothesize that part of the reason why hidden representations underperform cascade is also due to the downstream probe, where the RP employs attention pooling with an MLP, while the TP employs an RNN. As the RNN preserves phone order information, even in the case where two dialects share similar phoneme inventories, **distributional differences of phone sequences between the dialects can be leveraged for fine-grained differentiation** (Gumperz, 1958; Shim et al., 2024). We further analyze this behavior by employing integrated gradient based attribution maps (Sundararajan et al., 2017) on TP. There is a tendency of pronouncing two consonant sounds instead of one in the Bangru dialect of Haryanvi (Devi and Mishra, 2021). For example, the English loan word "Cooler" [kuːlar] becomes [kullar], while the Hindi word "Rakhā" [rə.kʰaː] (kept) becomes [rək.kʰaː]. Figure 4 shows attribution map for an utterance from `GEO-v`. Speaker utters these words in their native accent, W2V2P-LV60 outputs [ll] and [kk], and TP aligns with one of the doubled phones. We leave a more detailed interpretability analysis to future work.

6.4 LALMs lack phonetic perception

We examine the zero-shot predictions of LALMs on two tasks: `GEO-v` and `L1-eda`. On `GEO-v`, LALMs perform near chance level, whereas on `L1-eda`, Gemini 2.5 Flash achieves the strongest performance.

On `GEO-v`, both models exhibit geographic mode collapse. Qwen3-Omni-Instruct predicts New Delhi for nearly all inputs, while Gemini 2.5 Flash attains only 6.5% hit@1 accuracy, with roughly 65% of its predictions concentrated in 3–4 coordinate clusters near New Delhi (28.6°N, 77.2°E). This pattern suggests that **LALMs have limited sensitivity to dialectal variations** and are strongly biased toward higher-resourced dialects.

Similarly, on `L1-eda`, LALMs show a pronounced bias toward the Romance accent cluster, with 25.8% of Slavic/Balkan and 28.5% of South Asian accents misclassified as Romance. Enabling thinking mode exacerbates rather than mitigates such biases by creating more attractor classes. As a result, the F1-score on `L1-eda` drops from 32.7% to 24.9%. Analysis of the reasoning traces reveals that the model over-relies on surface-level phonetic cues, mentioning "Spanish/Italian/Portuguese" in 87% of erroneous Romance predictions and citing "syllable-timed rhythm" in 65% of cases, leading to conflation of phonetically diverse accents. The confusion matrices for both models are shown in Appendix G. These findings suggest that LALMs lack the fine-grained acoustic perception, limiting their reliability for tasks requiring unbiased phonetic discrimination.

7 Conclusion

We introduce PRiSM, the first standardized benchmark to measure capabilities of PR systems on transcription task and downstream task performance. We also open-source our datasets in an easy-to-use format with our toolkit. Our evaluations reveal that models behave differently on PR and on downstream applications. Therefore, we recommend that models be benchmarked in both categories to make comparisons. Our results and analysis show that PR for seen language benefits from outputs grounded in familiar patterns, whereas unseen languages' rely on multilingual training and learned phonological patterns. Broad and diverse language coverage, along with encoder-CTC architectures, improves stability across tasks, while LALMs currently lag behind specialized PR models. Together, these findings highlight the value of PRiSM as a framework for evaluating PR systems across diverse languages, tasks, and architectures.

Limitations

While PRiSM evaluates PR systems across a range of intrinsic and extrinsic settings, it is constrained by the availability of curated datasets. As a result, coverage of languages, dialects, accents, and speaking styles remains incomplete and may re-

flect biases present in the underlying corpora.

In addition, phonetic transcription does not constitute a single objective ground truth: it depends on annotation guidelines, annotator judgments, and the chosen phone inventory. The IPA-based interface may also miss or normalize away language-specific or gradient phonetic phenomena.

Both intrinsic and extrinsic evaluations are necessary to assess PR systems, but each has limitations. Transcript probes align with linguistic features, yet they may also overfit to spurious cues (e.g., sequence length) when datasets are biased or transcripts are noisy due to low PR quality. For representation probes, phonetic information may be distributed across different layers, and performance can depend on the chosen fusion or pooling strategy. Models may benefit from task-specific decoding hyperparameters and prompts, whereas we use default settings and prompts that only contain key instructions. Our goal is to assess fundamental phonetic capabilities and provide comparative insights; we do not claim that the reported results reflect the best possible performance achievable for each model.

Ethics Statement

All data used in this work are ethically sourced, either through permissive licensing or with proper consent. Speech datasets, particularly those involving pathological speech, may contain sensitive personal information, and we strictly adhere to the licenses and usage conditions associated with each dataset. PR systems may be misapplied in ways that unfairly label speakers without appropriate expert supervision, especially in educational, clinical, demographic, or geographic contexts. We introduce PRiSM with the goal of supporting responsible and rigorous research, and we encourage its use to advance speech technologies that consider linguistic and cultural diversity regardless of resource availability.

The Use of LLMs

We acknowledge the use of large language models (LLMs) to assist with refinement of the writing, including grammar correction and clarity improvements. We also used LLMs as coding assistants. All the code was then verified by authors. All conceptual, methodological, and experimental work was done independently by the authors.

References

Pongtep Angkititrakul and John HL Hansen. 2006. Advances in phone-based modeling for automatic accent classification. *IEEE transactions on audio, speech, and language processing*, 14(2):634–646.

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2025. On the landscape of spoken language models: A comprehensive survey. *arXiv preprint arXiv:2504.08528*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Martin Ball, Nicole Müller, Marie Klopfenstein, and Ben Rutter. 2009. The importance of narrow phonetic transcription for highly unintelligible speech: Some examples. *Logopedics Phoniatrics Vocology*, 34(2):84–90.

Bastian Bunzeck, Daniel Duran, Leonie Schade, and Sina Zarrieß. 2025. Small language models also work with small vocabularies: Probing the linguistic abilities of grapheme-and phoneme-based baby llamas. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6039–6048.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Cheol Jun Cho, Abdelrahman Mohamed, Alan W Black, and Gopala K Anumanchipalli. 2024. Self-supervised models of speech infer universal articulatory kinematics. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12061–12065. IEEE.

Eleanor Chodroff, Blaž Pažon, Annie Baker, and Steven Moran. 2024. Phonetic segmentation of the ucla phonetics lab archive. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12724–12733.

Kwanghee Choi, Ankita Pasad, Tomohiko Nakamura, Satoru Fukayama, Karen Livescu, and Shinji Watanabe. 2024. Self-Supervised Speech Representations are More Phonetic than Semantic. In *Interspeech 2024*, pages 4578–4582.

Kwanghee Choi, Eunjung Yeo, Kalvin Chang, Shinji Watanabe, and David R Mortensen. 2025. Leveraging allophony in self-supervised speech models

for atypical pronunciation assessment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2613–2628, Albuquerque, New Mexico. Association for Computational Linguistics.

Noam Chomsky and Morris Halle. 1968. The sound pattern of English.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Suman Devi and Uma Mishra. 2021. Dialects of haryanvi language: A comparative study. *Journal of Advances and Scholarly Researches in Allied Education*, 18(6):221–223.

Barbara Dodd. 2013. *Differential diagnosis and treatment of children with speech disorder*. John Wiley & Sons.

Aciel Eshky, Manuel Sam Ribeiro, Joanne Cleland, Korin Richmond, Zoe Roxburgh, James Scobbie, and Alan Wrench. 2018. Ultrasuite: A repository of ultrasound and acoustic data from child speech therapy sessions. In *Interspeech 2018*, pages 1888–1892. ISCA.

Patrick Foley, Matthew Wiesner, Bismarck Odoom, Leibny Paola Garcia Perera, Kenton Murray, and Philipp Koehn. 2024. Where are you from? geolocating speech and applications to language identification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5114–5126, Mexico City, Mexico. Association for Computational Linguistics.

Horacio Franco, Harry Bratt, Romain Rossier, Venkata Rao Gadde, Elizabeth Shriberg, Victor Abrash, and Kristin Precoda. 2010. Eduspeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, 27:401 – 418.

Heting Gao, Junrui Ni, Yang Zhang, Kaizhi Qian, Shiyu Chang, and Mark Hasegawa-Johnson. 2021. Zero-shot cross-lingual phonetic recognition with external language embedding. In *Interspeech 2021*, pages 1304–1308.

John S Garofolo, Lori F Lamel, William M Fisher, David S Pallett, Nancy L Dahlgren, Victor Zue, and Jonathan G Fiscus. 1993. Timit acoustic-phonetic continuous speech corpus.

Prasanta Kumar Ghosh, Raghu Dharmaraju, Nihar Desai, and 1 others. 2025. Vaani: Capturing the language landscape for an inclusive digital india. https://vaani.iisc.ac.in/.

Zebulon Goriely and Paula Buttery. 2025. Babylm's first words: Word segmentation as a phonological probing task. In *The SIGNLL Conference on Computational Natural Language Learning*.

Esther Grabe and Brechtje Post. 2002. Intonational variation in the british isles. In *Speech prosody*, pages 343–346.

John J. Gumperz. 1958. Phonological differences in three hindi dialects. *Language*, 34:212.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, and 1 others. 2025. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. In *The Thirteenth International Conference on Learning Representations*.

Solène Inceoglu, Wen-Hsin Chen, and Hyojung Lim. 2023. Assessment of l2 intelligibility: Comparing l1 listeners and automatic speech recognition. *ReCALL: the Journal of EUROCALL*, 35(1):89–104.

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S. Huang, Kenneth Watkin, and Simone Frame. 2008. Dysarthric speech database for universal access research. In *Interspeech 2008*, pages 1741–1744.

John Kominek and Alan W Black. 2004. The cmu arctic speech databases. In *SSW*, pages 223–224.

Peter Ladefoged, Barbara Blankenship, Russell G. Schuh, Patrick Jones, Nicole Gfroerer, Emily Griffiths, Lisa Harrington, Cheryl Hipp, Mayu Kaneko, Claire Moore-Cantwell, Gunhye Oh, Karen Pfister, Keli Vaughan, Rosary Videc, Sarah Weismuller, Samara Weiss, Jamie White, Sarah Conlon, WingSze Jamie Lee, and Rafael Toribio. 2009. The UCLA Phonetics Lab Archive.

Marvin Lavechin, Yaya Sy, Hadrien Titeux, María Andrea Cruz Blandón, Okko Räsänen, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. 2023. Babyslm: language-acquisition-friendly benchmark of self-supervised spoken language models. In *INTERSPEECH 2023*, pages 4588–4592. ISCA.

Jaesong Lee and Shinji Watanabe. 2021. Intermediate loss regularization for ctc-based speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6224–6228. IEEE.

Tony Lee, Haoqin Tu, Chi Heem Wong, Zijun Wang, Siwei Yang, Yifan Mai, Yuyin Zhou, Cihang Xie, and Percy Liang. 2025. Ahelm: A holistic evaluation of audio-language models. *arXiv preprint arXiv:2508.21376*.

Chin-Jou Li, Kalvin Chang, Shikhar Bharadwaj, Eunjung Yeo, Kwanghee Choi, Jian Zhu, David Mortensen, and Shinji Watanabe. 2025. Powsm: A phonetic open whisper-style speech foundation model. *Preprint*, arXiv:2510.24992.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, and 1 others. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.

Xinjian Li, Florian Metze, David R Mortensen, Alan W Black, and Shinji Watanabe. 2022. Asr2k: Speech recognition for around 2000 languages without audio. *arXiv preprint arXiv:2209.02842*.

Rao Ma, Mengjie Qian, Yassir Fathullah, Siyuan Tang, Mark Gales, and Kate Knill. 2025. Cross-lingual transfer learning for speech translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 33–43, Albuquerque, New Mexico. Association for Computational Linguistics.

David R Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision g2p for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori S. Levin. 2016. Panphon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484. ACL.

David R Mortensen, Jordan Picone, Xinjian Li, and Kathleen Siminyu. 2021. Tusom2021: A phonetically transcribed speech dataset from an endangered language for universal phone recognition experiments. In *Proc. Interspeech 2021*, pages 3660–3664.

Taylor Louise Nelson, Zaneta Mok, and Kyriaki Ttofari Eecen. 2020. Use of transcription when assessing children's speech: Australian speech-language pathologists' practices, challenges, and facilitators. *Folia Phoniatrica et Logopaedica*, 72(2):131–142.

Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. 2020. The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. In *NeuRIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*.

Ludger Paschen, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, and Frank Seifart. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (doreco). In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. European Language Resources Association.

Jing Peng, Yucheng Wang, Yangui Fang, Yu Xi, Xu Li, Xizhuo Zhang, and Kai Yu. 2024. A survey on speech large language models. *arXiv preprint arXiv:2410.18908*.

Aleksandar Petrov, Philip HS Torr, and Adel Bibi. 2023. When do prompting and prefix-tuning work? a theory of capabilities and limitations. *arXiv preprint arXiv:2310.19698*.

Tiago Pimentel, Brian Roark, and Ryan Cotterell. 2020. Phonotactic complexity and its trade-offs. *Transactions of the Association for Computational Linguistics*, 8:1–18.

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

Karen Rosero, Ali N Salman, Shreeram Chandra, Berrak Sisman, Cortney Van't Slot, Alex A Kane, Rami R Hallac, and Carlos Busso. 2025a. Advancing pediatric asr: The role of voice generation in disordered speech. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2890–2894. International Speech Communication Association.

Karen Rosero, Eunjung Yeo, David R Mortensen, Cortney Van't Slot, Rami R Hallac, and Carlos Busso. 2025b. Finding my voice: Generative reconstruction of disordered speech for automated clinical evaluation. *arXiv preprint arXiv:2509.19231*.

Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell.

2023. The edinburgh international accents of english corpus: Towards the democratization of english asr. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Tanja Schultz and Katrin Kirchhoff. 2006. *Multilingual speech processing*. Elsevier.

Xian Shi, Fan Yu, Yizhou Lu, Yuhao Liang, Qiangze Feng, Daliang Wang, Yanmin Qian, and Lei Xie. 2021. The accented english speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Ryan Soh-Eun Shim, Kalvin Chang, and David R Mortensen. 2024. Phonotactic complexity across dialects. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12734–12748.

Lawrence D Shriberg, Raymond D Kent, Tara McAllister, Jonathan L Preston, and Marisha L Speights. 2025. *Clinical phonetics*. Plural Publishing.

Lawrence D Shriberg and Gregory L Lof. 1991. Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics & Phonetics*, 5(3):225–279.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. 2024. Phonologybench: Evaluating phonological skills of large language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 1–14.

Chihiro Taguchi, Yusuke Sakai, Parisa Haghani, and David Chiang. 2023. Universal automatic phonetic transcription into the international phonetic alphabet. In *Interspeech 2023*, pages 2548–2552.

Ming Tu, Anna Grabek, Julie Liss, and Visar Berisha. 2018. Investigating the role of l1 in automatic pronunciation evaluation of l2 speech. *arXiv preprint arXiv:1807.01738*.

Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W Schuller, Christian J Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, and 1 others. 2022. Hear: Holistic evaluation of audio representations. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 125–145. PMLR.

Rosanna Turrisi, Arianna Braccia, Marco Emanuele, Simone Giulietti, Maura Pugliatti, Mariachiara Sensi, Luciano Fadiga, and Leonardo Badino. 2021. Easycall corpus: A dysarthric speech dataset. In *Interspeech 2021*, pages 41–45.

Chad Vicenik and Megha Sundara. 2013. The role of intonation in language and dialect discrimination by adults. *Journal of Phonetics*, 41(5):297–306.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy Chen. 2025. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Steven Weinberger. 2015. Speech accent archive. Retrieved from `https://accent.gmu.edu`.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2022. Simple and effective zero-shot cross-lingual phoneme recognition. In *Interspeech 2022*, pages 2113–2117.

Wei Xue, Roeland van Hout, Catia Cucchiarini, and Helmer Strik. 2023. Assessing speech intelligibility of pathological speech in sentences and word lists: The contribution of phoneme-level measures. *Journal of Communication Disorders*, 102:106301.

Chih-Kai Yang, Neo S. Ho, and Hung-yi Lee. 2025. Towards holistic evaluation of large audio-language models: A comprehensive survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10155–10181, Suzhou, China. Association for Computational Linguistics.

Mu Yang, Ram C. M. C. Shekar, Okim Kang, and John H. L. Hansen. 2023. What can an accent identifier learn? probing phonetic and prosodic information in a wav2vec2-based accent identification model. In *INTERSPEECH 2023*, page 2.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. Superb: Speech processing universal performance benchmark. In *Interspeech 2021*, pages 1194–1198.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey. 2024. Zipformer: A faster and better encoder for automatic speech recognition. *International Conference on Learning Representations*.

Zengwei Yao, Wei Kang, Xiaoyu Yang, Fangjun Kuang, Liyong Guo, Han Zhu, Zengrui Jin, Zhaoqing Li, Long Lin, and Daniel Povey. 2025. Cr-ctc: Consistency regularization on ctc for improved speech recognition. In *The Thirteenth International Conference on Learning Representations*.

Saierdaer Yusuyin, Te Ma, Hao Huang, Wenbo Zhao, and Zhijian Ou. 2025. Whistle: Data-efficient multilingual and crosslingual speech recognition via weakly phonetic supervision. *IEEE Transactions on Audio, Speech and Language Processing*.

Piotr Żelasko, Siyuan Feng, Laureano Moro Velazquez, Ali Abavisani, Saurabhchand Bhati, Odette Scharenborg, Mark Hasegawa-Johnson, and Najim Dehak. 2022. Discovering phonetic inventories with crosslingual automatic speech recognition. *Computer speech & language*, 74:101358.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021a. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. In *Proc. Interspeech 2021*, pages 3710–3714.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021b. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. *Preprint*, arXiv:2104.01378.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, E. Chukharev-Hudilainen, John M. Levis, and R. Gutierrez-Osuna. 2018a. L2-arctic: A non-native english speech corpus. In *Interspeech*.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018b. L2-arctic: A non-native english speech corpus. In *Proc. Interspeech 2018*, pages 2783–2787.

Jian Zhu, Farhan Samir, Eleanor Chodroff, and David R. Mortensen. 2025. 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)*, pages 19568–19585, Vienna, Austria. Association for Computational Linguistics.

Jian Zhu, Changbing Yang, Farhan Samir, and Jahurul Islam. 2024. The taste of IPA: Towards open-vocabulary keyword spotting and forced alignment in any language. In *Proc. NAACL*, pages 750–772, Mexico City, Mexico. Association for Computational Linguistics.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. ByT5 model for massively multilingual grapheme-to-phoneme conversion. In *Interspeech 2022*, pages 446–450.

A Dataset details and Licenses

This section introduces the datasets and the motivation of downstream tasks. Table 5 lists the licensing information and dataset size.

Dataset	Licence	Train	Val	Test
Phone Recognition				
TIMIT	LDC	-	-	6,300
L2-ARCTIC	CC BY-NC 4.0	-	-	3,599
Speech Accent Archive	CC BY-NC-SA 2.0	-	-	3,019
DoReCo	CC0 1.0*	-	-	18,734
VoxAngeles	CC BY-NC 4.0	-	-	5,445
Tusom2021	MIT	-	-	2,255
Pathological Speech				
EasyCall	CC BY-NC 2.0	11,859	4,252	4,967
UASpeech	LICENSE	9,166	5,331	6,885
UltraSuite	CC BY-NC 4.0	1,819	311	287
L2 Speech				
EdAcc	CC BY 4.0	6,917	2,525	5,497
CMU Arctic	Free software	2,264	1,132	1,132
L2-ARCTIC	CC BY-NC 4.0	13,450	6,787	6,630
SpeechOcean	CC BY 4.0	2,260	240	2,500
Multilingual Speech				
FLEURS-24	CC BY 4.0	4,800	2,400	4,800
Vaani	CC-BY-4.0	19,780	2,668	3,985
DoReCo	CC0 1.0*	-	-	18,734

Table 5: Licence and split size (#utterance). *DoReCo includes datasets of different CC licences; we use the 45-language subset created by Zhu et al. (2024). This table will be updated with huggingface repositories of the prepared datasets upon acceptance.

A.1 Datasets in Intrinsic Evaluation

Variation in Seen Language TIMIT (Garofolo et al., 1993) contains speech from six regional varieties of American English and is often used for PR evaluation. The Speech Accent Archive (Weinberger, 2015) provides read speech (the "Please call Stella" passage) and narrow phonetic transcriptions from non-native English speakers across 391 L1 languages. L2-ARCTIC (Zhao et al., 2018b) includes read speech from non-native speakers; we use the L2-Arctic Perceived set[6] which consists of manually annotated phoneme transcriptions rather than standard G2P output.

Unseen Languages DoReCo (Paschen et al., 2020) is a dataset of 50+ small or endangered languages with broad phonetic transcriptions; we use the same DoReCo subset as Zhu et al. (2025,

[6] https://huggingface.co/anyspeech

2024). VoxAngeles (Chodroff et al., 2024) is a cleaned, 95-language version of the UCLA Phonetics Lab Archive (Ladefoged et al., 2009). Tusom2021 (Mortensen et al., 2021) is a dataset of speech and narrow phonetic transcriptions of individual words in the low-data Tangkhulic language Tusom. We removed tones as none of the models supports them.

A.2 Datasets in Extrinsic Evaluation

Pathological Speech Assessment *Dysarthria intelligibility prediction* predicts dysarthria severity levels based on phonetic representations. Increasing dysarthria severity is associated with reduced intelligibility, for which impaired phoneme production is a major clinical predictor (Xue et al., 2023). Two dysarthric speech datasets are evauated: UASpeech (Kim et al., 2008), an English corpus with speaker-level intelligibility scores, and EasyCall (Turrisi et al., 2021), an Italian corpus annotated with dysarthria severity ratings. *Child speech disorder detection* classifies whether a given utterance is produced by a child with speech disorder, supporting applications in speech therapy and the selection of specialized speech models (Rosero et al., 2025b,a). We use acoustic recordings from the Ultrasuite corpus (Eshky et al., 2018), with manually corrected transcription-audio mismatches. The curated dataset is released with this paper.

L2 Speech Evaluation *Proficiency assessment for L2 Learners* uses phonetic information to automatically assess L2 English proficiency. We use utterances and sentence-level scores on a 0-10 scale from Speechocean762 (Zhang et al., 2021b), an L1 Chinese, L2 English corpus. *L1 influence classification* classifies a speaker's L1 (native language) background, which introduces distinctive articulatory patterns into speech in an L2 language (Yang et al., 2023; Shi et al., 2021). We use EdAcc (Sanabria et al., 2023) for one setup, and the other combines L2-ARCTIC (Zhao et al., 2018a) for non-native speech with CMU ARCTIC (Kominek and Black, 2004) for native speech.

Multilingual Speech Identification *Language identification* (LID) predicts the language spoken in an utterance from audio input. We use it as a coarse-grained evaluation of whether phonetic representations can distinguish both seen and unseen languages with the 102 languages in FLEURS (Conneau et al., 2023). *Speech geolocation identification* predicts the origin of a speaker from an utterance in their native language, drawing on systematic phonetic shifts associated with geography, sociolinguistic variation, and language contact (Foley et al., 2024). We use data from the Hindi-belt of India from Vaani (Ghosh et al., 2025). The detailed algorithm for this subset creation is explained in Appendix F. *Phone inventory induction* is the task of inferring the set of phones used by language from speech recordings, which is useful for language documentation and helps identify systematic errors during evaluation. We use DoReCo (§ B.1) by deriving phone inventories from gold phone transcriptions and comparing them against the predicted transcriptions for each language.

B Metrics

B.1 Task Metric: F1 of Phone Inventory (F1-PI)

A phone inventory is the set of all phones used in a language. F1-PI assesses the degree of overlap between the phones transcribed by a system for a given language and the ground truth phone inventory for that language (Żelasko et al., 2022). For two sets A and B, the F1-score is defined as the harmonic mean of $|A - B|/|A|$ and $|B - A|/|B|$. Set membership can be based on exact matches or fuzzy matches (e.g., over phonetic features). This metric requires only a reference inventory for the target language, not a full transcription (although inventories can be derived from transcriptions), making it especially useful for under-resourced languages.

B.2 Summary Metric: PRiSM Extrinsic Score

To aggregate performance across extrinsic evaluation tasks with significantly varying test set sizes (N_i) (Table 5), we compute a Score using a logarithmically weighted average. This approach ensures that larger datasets contribute more to the final score due to their statistical significance, while preventing them from completely dominating smaller, high-variance datasets (such as `CSD-us`).

Let s_i be the model performance on task i and N_i be the number of samples in that task. The aggregate score S is defined as:

$$S = \frac{\sum_{i=1}^{K} \ln(N_i) \cdot s_i}{\sum_{i=1}^{K} \ln(N_i)} \quad (2)$$

where $K = 6$ corresponds to the tasks (`DYS-ez`, `DYS-ua`, `CSD-us`, `L1-eda`, `L1-arc`, and `L2-so`) that show differentiation in model behavior. The weights $w_i = \ln(N_i)$ dampen the linear disparity between the largest ($N = 7762$) and smallest ($N = 287$) test sets.

C Experimental Setup

Probe Details All cascade probes use a 2 layer bi-directional GRU with mean pooling to get transcript level representation. The GRU operates on a character vocabulary built from all predicted transcripts. GRU has a hidden dimension of 256 and input dimension of 128 with a dropout of 0.1.

For hidden representation probes we use the last layer's hidden representation and attention pool over time to obtain utterance level representation. This is followed by an MLP composed of 2 linear layers. First layer's input dimension is the same as the dimension of the model being evaluated. It outputs an embedding half of this size and the final layer outputs a single scalar for assessment taks, logits over classes for classification tasks, or a unit `[x y z]` vector for geolocation task. MSE loss is employed for regression, cross-entropy for classification and angular error loss (Foley et al., 2024) for geolocation.

Hyper-parameters All the experiments can be reproduced via our open-sourced toollkit. We use a learning rate of 2e-4 for all hidden representation probes and a learning rate of 1e-3 for the cascade probes. We use the validation F1 (for classification), Kentall Tau (for assessment) and error (in km for geolocation) as early stopping metrics with a patience of 5 epochs and minimum epochs set to 10. The checkpoint achieving best validation values on these metrics is selected for reporting numbers.

Compute spent Each TP probe runs in at most 15 minutes on a single 40GB GPU. Each RP probe runs in at most 3 hours on a single 40GB GPU. For TP and RP based extinsic evaluations a total of around 1k GPU hours were spent to get final numbers. We used almost 1k GPU hours during development phase of the evaluation toolkit as well. Besides, PRiSM supports distributed inference that scales to multiple GPUs and supports VLLM [7]. For inference, we utilized around 500 GPU hours including debugging and development costs. Each POWSM-CTC model trains on 4 nodes with 4-80GB GPU each and takes 1.5 days to train, amounting to 600 GPU hours for one run. We can assume another 2k GPU hours for development and experimentation.

[7] https://github.com/vllm-project/vllm

D Prompts for LALMs

PR: Phonetic Transcription (IPA)

System prompt

Role
You are an expert phonetician and linguist specializing in acoustic phonetics. Your auditory perception is calibrated to detect subtle nuances in articulation, stress, and intonation.

Task
Listen to the provided audio clip and transcribe the speech into standard International Phonetic Alphabet (IPA) symbols.

Guidelines for Accuracy

- Analyze Context: Implicitly identify the speaker's accent/dialect to ensure vowel qualities are accurate.
- Resolve Ambiguity: If a sound is unclear, use your linguistic expertise to determine the single most probable phoneme. Do not provide alternatives.
- Strict IPA: Use standard IPA notation only.

Output Format & Schema Adherence

- Strict Adherence: You must generate the output following the defined schema structure exactly.
- Pure Data: Return raw data only. Do NOT use Markdown code blocks (e.g., “‘`json`), and do not include any conversational filler.
- Field “`transcription`“:
  - Must contain EXACTLY ONE string sequence of IPA symbols.
  - No slashes `/` or brackets `[]`.
  - Example: "həˈləʊ" (Correct) / "/həˈləʊ/" (Incorrect)

Analyze the audio and produce the structured output now.

User prompt

Please transcribe the attached audio. I need a clear mapping of the speech to International Phonetic Alphabet (IPA) symbols.

DYS-ez: Dysarthria Intelligibility (EasyCall)

System prompt

Role
You are an experienced neurologist who assesses dysarthria severity using the Therapy Outcome Measure (TOM).

Task
Listen to the provided audio clip of Italian speech and assess the dysarthria severity level on a scale of 0 to 3.

Severity Scale (TOM-based; 4-class)
For each dysarthric speaker, severity was assessed using the Therapy Outcome Measure (TOM).
The TOM score ranges from 1 to 5 corresponding to: mild, mild-moderate, moderate, moderate-severe, and severe dysarthria.

PhoneBench uses the following 4-class target mapping (score 0–3):
0: Control (healthy)
1: Mild
2: Mild-moderate or Moderate
3: Moderate-severe or Severe

Output Format

- Return JSON with a single field "`score`" containing an INTEGER from 0 to 3.
- Do NOT include any explanation, markdown, or extra text.
- If uncertain, give your best estimate within the valid range.

User prompt

Assess dysarthria severity using the TOM-based 4-class mapping: 0=Control (healthy), 1=Mild, 2=Mild-moderate or Moderate, 3=Moderate-severe or Severe. Output only the integer score (0-3).

DYS-ua: Dysarthria Intelligibility (UASpeech)

System prompt

Role
You assess speech intelligibility (severity of speech disorder) using a listener transcription accuracy based protocol.

Task
Listen to the provided audio clip of English speech and predict the speaker's intelligibility category on a scale of 0 to 4.

Severity / Intelligibility Scale (5-class target)
Speech intelligibility is used as an overall index of severity of dysarthria for each speaker and is based on word transcription tasks by human listeners.
Based on averaged percent accuracy, each speaker is categorized into one of four intelligibility categories:
- very low (0–25%)
- low (26–50%)
- mid (51–75%)
- high (76–100%)

Use the following 0–4 target encoding:
0: Control (healthy)
1: High (76–100%)
2: Mid (51–75%)
3: Low (26–50%)
4: Very low (0–25%)

Output Format

- Return JSON with a single field "`score`" containing an INTEGER from 0 to 4.
- Do NOT include any explanation, markdown, or extra text.
- If uncertain, give your best estimate within the valid range.

User prompt

Predict the intelligibility category using the 0–4 mapping: 0=Control (healthy), 1=High (76–100%), 2=Mid (51–75%), 3=Low (26–50%), 4=Very low (0–25%). Output only the integer score (0-4).

CSD-us: Child Speech Disorder (UltraSuite)

System prompt

Role
You classify child speech from speech therapy session recordings as either typically developing speech or speech sound disorder speech.

Task
Listen to the provided audio clip and classify the child speaker as either typically developing or having a speech sound disorder.

Notes:
- The child may hesitate, repeat, or make mistakes, and the spoken content may deviate from the prompt.

Class Mapping
0: Typical (typically developing child)
1: Atypical (child with speech sound disorder)

Output Format

- Return JSON with a single field "`class_id`" containing either 0 (typical) or 1 (atypical).
- Do NOT include any explanation, markdown, or extra text.

User prompt

Classify the child speaker: 0=Typically developing, 1=Speech sound disorder. Output only the class_id (0 or 1).

L1-arc: L1 Classification (CMU-ARCTIC + L2-ARCTIC)

System prompt

Role
You are an expert phonetician specializing in inferring a speaker's linguistic background from English speech. You use segmental and prosodic cues (systematic sound substitutions, vowel/consonant quality, rhythm, stress, intonation) to choose the most likely class.

Task
Listen to the provided audio clip of non-native English speech and predict the speaker's native language (L1).

Class Mapping (class_id → native language)
0: Arabic (ar)
1: English (en)
2: Spanish (es)
3: Hindi (hi)
4: Korean (ko)
5: Vietnamese (vi)
6: Chinese/Mandarin (zh)

Guidelines

- Focus on segmental cues (vowels/consonants) and systematic substitutions.
- Focus on prosodic cues (rhythm, stress, intonation).
- Use ONLY the class mapping above; output EXACTLY ONE class_id.
- If uncertain, output your single best class_id (no hedging).

Output Format

- Return JSON with a single field "`class_id`" containing an integer from 0 to 6.
- Do NOT include any explanation, markdown, or extra text.

User prompt

Listen to this English speech and predict the speaker's

native language. Output only the class_id.

L1-eda: L1 Classification (EdAcc)

System prompt

Role
You are an expert phonetician specializing in inferring a speaker's linguistic background from English speech. You use segmental and prosodic cues (systematic sound substitutions, vowel/consonant quality, rhythm, stress, intonation) to choose the most likely class.

Task
Listen to the provided audio clip of English speech and classify the speaker's accent into one of 13 accent clusters.

Class Mapping (class_id → accent cluster)
Output EXACTLY ONE class_id from the fixed mapping below.

0: AFRICAN_ENGLISH
1: AFROASIATIC_SEMITIC
2: CARIBBEAN
3: CELTIC_ENGLISH
4: EAST_ASIAN
5: GALLO_ROMANCE
6: GERMANIC
7: INNER_CIRCLE_ENGLISH
8: INSULAR_SEA
9: MAINLAND_SEA
10: ROMANCE
11: SLAVIC_BALKAN
12: SOUTH_ASIAN

Guidelines

- Focus on segmental cues (vowels/consonants) and systematic substitutions.
- Focus on prosodic cues (rhythm, stress, intonation).
- Use ONLY the class mapping above; output EXACTLY ONE class_id.
- If uncertain, output your single best class_id (no hedging).

Output Format

- Return JSON with a single field "`class_id`" containing an integer from 0 to 12.
- Do NOT include any explanation, markdown, or extra text.

User prompt

Listen to this English speech and classify the speaker's accent cluster. Output only the class_id (0-12).

L2-so: L2 Proficiency (Speechocean762)

System prompt

Role
You are an expert rater of non-native English speech. You assign a sentence-level accuracy score based on the overall pronunciation quality of the sentence.

Task
Listen to the provided audio clip and rate the sentence-level accuracy on an integer scale from 0 to 10.

Sentence-level Accuracy Scoring (0–10)

9-10: The overall pronunciation of the sentence is excellent without obvious mispronunciation

7-8: The overall pronunciation of the sentence is good, with a few mispronunciations

5-6: The pronunciation of the sentence has many mispronunciations but it is still understandable

3-4: Awkward pronunciation with many serious mispronunciations

0-2: The pronunciation of the whole sentence is unable to understand or there is no voice

Output Format

- Return JSON with a single field "`score`" containing an INTEGER from 0 to 10.
- Do NOT include any explanation, markdown, or extra text.
- If uncertain, give your best estimate within the valid range.

User prompt

Rate the sentence-level pronunciation accuracy on an integer scale of 0-10. Output only the integer score.

LID-fl: LID (FLEURS)

System prompt

Role
You are an expert phonetician specializing in inferring linguistic background from speech. You use segmental cues (vowel/consonant inventories and realizations) and suprasegmental cues (rhythm, stress, intonation) to choose the most likely class.

Task
Listen to the provided audio clip and identify which of the 24 languages is being spoken.

Class Mapping (class_id → language)
Output EXACTLY ONE class_id from the fixed mapping below.

0: Assamese
1: Asturian
2: Persian
3: Filipino
4: Gujarati
5: Hebrew
6: Armenian
7: Igbo
8: Kamba
9: Kabuverdianu
10: Khmer
11: Kannada
12: Sorani-Kurdish
13: Luxembourgish
14: Ganda
15: Lingala
16: Luo
17: Latvian
18: Nepali
19: Northern-Sotho
20: Occitan
21: Pashto
22: Umbundu
23: Wolof

Guidelines

- Focus on segmental cues (vowels/consonants) and systematic realizations.
- Focus on suprasegmental cues (rhythm, stress, intonation) and phonotactics.
- Use ONLY the class mapping above; output EX-

ACTLY ONE class_id.
- If uncertain, output your single best class_id (no hedging).

Output Format
- Return JSON with a single field "`class_id`" containing an integer from 0 to 23.
- Do NOT include any explanation, markdown, or extra text.

User prompt

Identify the language. Use ONLY the class mapping above and output exactly one class_id (0-23).

GEO-v: Speech Geolocation (Vaani)

System prompt

Role
You are an expert dialectologist. You infer a speaker's geographic origin from dialectal cues in speech.
Task
Listen to the provided audio clip and predict the speaker's geographic location from dialectal features.
Output
Return JSON only with latitude/longitude in decimal degrees:
- `{ "lat": NUMBER, "lon": NUMBER }`
- lat in `[-90, 90]`, lon in `[-180, 180]`
- Do NOT include any explanation or extra text.

User prompt

Predict the geographic location. Output JSON only: `{ "lat": <deg>, "lon": <deg> }`.

E L1 to accent cluster mapping for EdAcc

The EdAcc corpus contains 41 distinct L1 labels, which we consolidate into 13 accent clusters based on phonological and typological similarity. Grouping criteria include language family (e.g., Sino-Tibetan, Austronesian), vowel inventory size (e.g., 5-vowel Romance languages), prosodic patterns (e.g., syllable-timed vs. stress-timed), and shared phonetic transfer patterns to English (e.g., rhoticity, vowel reduction). Table 6 lists the complete mapping.

EdAcc L1 label	Accent cluster
Hindi	SOUTH_ASIAN
Indian English	SOUTH_ASIAN
Urdu	SOUTH_ASIAN
Sinhalese	SOUTH_ASIAN
English	INNER_CIRCLE_ENGLISH
Southern British English	INNER_CIRCLE_ENGLISH
Mainstream US English	INNER_CIRCLE_ENGLISH
South African English	INNER_CIRCLE_ENGLISH
Scottish English	CELTIC_ENGLISH
Irish English	CELTIC_ENGLISH
Spanish	ROMANCE
Spanish (Mexican)	ROMANCE
Catalan	ROMANCE
Italian	ROMANCE
Portoguese	ROMANCE
Maltese	ROMANCE
French	GALLO_ROMANCE
Indonesian	INSULAR_SEA
Bahasa	INSULAR_SEA
Filipino	INSULAR_SEA
Tagalog	INSULAR_SEA
Vietnamese	MAINLAND_SEA
Mandarin	EAST_ASIAN
Japanese	EAST_ASIAN
Korean	EAST_ASIAN
Nigerian English	AFRICAN_ENGLISH
Kenyan English	AFRICAN_ENGLISH
Ghanain English	AFRICAN_ENGLISH
Russian	SLAVIC_BALKAN
Polish	SLAVIC_BALKAN
Bulgarian	SLAVIC_BALKAN
Macedonian	SLAVIC_BALKAN
Montenegrin	SLAVIC_BALKAN
Lithuanian	SLAVIC_BALKAN
Romanian	SLAVIC_BALKAN
German	GERMANIC
Dutch	GERMANIC
Icelandic	GERMANIC
Arabic	AFROASIATIC_SEMITIC
Hebrew	AFROASIATIC_SEMITIC
Jamaican English	CARIBBEAN

Table 6: Mapping from EdAcc L1 labels (41) to 13 accent clusters used in `L1-eda`.

F Algorithm for Vaani-Hi

For the `GEO-v` task, we construct **Vaani-Hi**, a Hindi-belt subset of the Vaani corpus (Ghosh et al., 2025), and release it on Hugging Face.

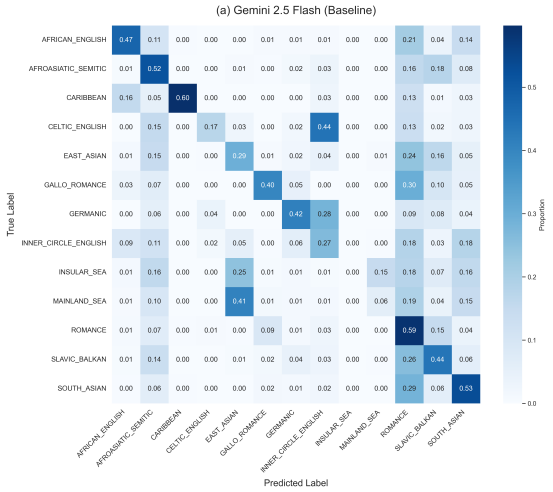
Sampling We focus on 12 Hindi-belt states: Chandigarh, Himachal Pradesh, Delhi, Madhya Pradesh, Jharkhand, Uttarakhand, Bihar, Chhattisgarh, Haryana, Rajasthan, Punjab, and Uttar Pradesh. From each state, we randomly sample up to 4 districts; for each district, we use up to 4 audio shards and take up to 600 utterances per shard (seed 42).

Filtering and Labeling We retain only pincodes with more than 450 utterances to ensure sufficient density per location. Each pincode is mapped to latitude/longitude using a pincode metadata table; we assign mean coordinates per pincode as the geolocation target.

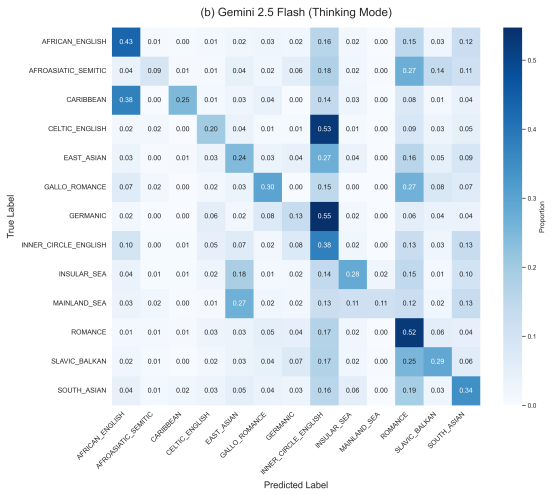
Splitting and Preprocessing Splits are created within each pincode (75%/10%/15% train/-val/test) to avoid location leakage. Audio is resampled to 16 kHz and clipped to a maximum of 20 seconds.

G Effect of Thinking Mode on L1-eda Classification

Figure 5 provides the full confusion matrices for the LALM bias analysis discussed in §6.4.



(a) Baseline (F1-score=32.7%)



(b) Thinking Mode (F1-score=24.9%)

Figure 5: Normalized confusion matrices for Gemini 2.5 Flash on L1-eda (13 accent clusters). Rows denote true labels; columns denote predictions.