POWSM : A Phonetic Open Whisper-Style Speech Foundation Model

Anonymous ACL submission

Abstract

Recent advances in spoken language processing have led to substantial progress in phonetic tasks such as automatic speech recognition (ASR), phone recognition (PR), grapheme-to-phoneme conversion (G2P), and phoneme-to-grapheme conversion (P2G). Despite their conceptual similarity, these tasks have largely been studied in isolation, each relying on task-specific architectures and datasets. In this paper, we introduce POWSM (Phonetic Open Whisper-style Speech Model), the first unified framework capable of jointly performing multiple phone-related tasks. POWSM enables seamless conversion between audio, text (graphemes), and phones, opening up new possibilities for universal and low-resource speech processing. Our model outperforms or matches specialized PR models of similar size (Wav2Vec2Phoneme and ZIPA) while jointly supporting G2P, P2G, and ASR. Our training data, models, and code will be released to foster open science.

1 Introduction

Phones are the smallest units of sound in speech. Unlike graphemes, phones are shared across languages and usually represented using the International Phonetic Alphabet (IPA) (International Phonetic Association, 1999), a unified transcription standard for all languages. By providing a consistent representation of speech across languages, phone-level modeling allows fine-grained analysis and cross-lingual generalization, enabling tasks like atypical speech analysis (*e.g.*, L2 speech (Li et al., 2016; Inceoglu et al., 2023) and pathological speech (Choi et al., 2025; Li et al., 2025)), endangered language documentation (He et al., 2024), code-switched text-to-speech (Zhou et al., 2020), and cross-lingual transfer in speech-to-text (Pratap et al., 2024; Magoshi et al., 2025).

Four key phone-related tasks underpin phonetic spoken language processing: automatic speech recognition (ASR), phone recognition (PR), grapheme-to-phoneme conversion (G2P), and phoneme-to-grapheme conversion (P2G). *ASR* learns implicit phonetic representations (Belinkov and Glass, 2017), while *PR* offers explicit phone-level supervision. *G2P* and *P2G* bridge orthographic and phonetic spaces. Collectively, these tasks interact through shared phonetic representations, each addressing a different aspect of the relationship between audio, phones, phonemes, and graphemes.

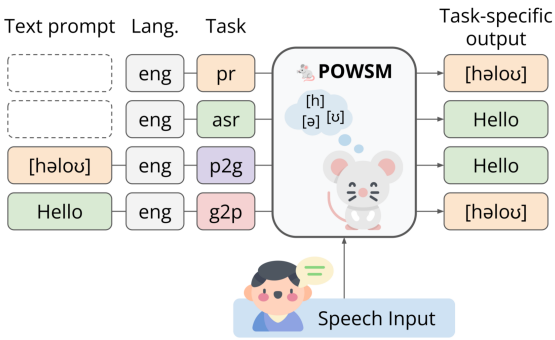

Figure 1: POWSM is the first phonetic foundation model that can perform four phone-related tasks: Phone Recognition (PR), Automatic Speech Recognition (ASR), audio-guided grapheme-to-phoneme conversion (G2P), and audio-guided phoneme-to-grapheme conversion (P2G).

Despite their conceptual similarity, these tasks have traditionally been developed in isolation, using task-specific architectures and datasets. Such systems are optimized for specific input-output mappings and cannot be easily extended to other phonetic tasks. This fragmentation has hindered the development of general-purpose models for phonetic processing, necessitating a unified phonetic foundation model that can perform multiple phone-related tasks within a single, general framework for speech processing.

To bridge this gap, we propose 🐁POWSM , a phonetic foundation model capable of performing four core phone-related tasks — PR, ASR, audio-guided G2P, and audio-guided P2G — within one unified architecture (Figure 1). To construct this framework, we reformulate standard ASR datasets (Zhu et al., 2025) into four task-specific formats, allowing the model to learn consistent mappings across audio, phoneme, and grapheme representations. In addition, 🐁POWSM adopts an attention-based encoder-decoder (AED) architecture, following the design of large-scale speech foundation models such as Whisper (Radford et al., 2023) and OWSM (Peng et al., 2023).

Empirically, 🐁POWSM outperforms previous PR models on both in-domain data and out-of-domain languages, achieves low-resource ASR performance comparable to web-scale multilingual foundation models, and can act as speech-grounded P2G and G2P across more than 70 languages. 🐁POWSM offers a new unified paradigm for phone-level modeling, paving the way for inclusive and globally accessible speech technologies that transcend language boundaries and resource disparities.

To summarize, our main contributions are:

- We provide 🐁POWSM , a large-scale foundation model that achieves state-of-the-art PR performance, and is capable of performing multiple fundamental phone-related tasks. Our model enables seamless conversion between speech, text (graphemes/orthography), and phones.
- We thoroughly analyze 🐁POWSM to understand the interaction between multiple tasks, architecture components, and losses.
- We fully open-source all our data preparation and evaluation scripts, model checkpoints and code to foster open science.

2 Related Work

Speech foundation models Recent speech foundation models such as Whisper (Radford et al., 2023) and OWSM (Peng et al., 2023, 2024) have driven progress in large-scale multilingual ASR and speech translation, but they do not explicitly address phoneme recognition or articulatory-level supervision. Subsequent work (Yusuyin et al., 2025; Fu et al., 2025) showed that incorporating phoneme-level objectives improves ASR for low-resource and long-tailed settings, while outputting phonemes as an intermediate benefited speech translation (Gállego et al., 2025). 🐁POWSM extends this line of work by being the first open foundation model jointly trained on phone recognition and related tasks, integrating multilinguality, phonetic supervision, and multi-task scalability within one framework.

Phone recognition Prior work in multilingual phone recognition can broadly be categorized into (1) language-specific models (Gao et al., 2021) that rely on explicit phoneme (Xu et al., 2022) or allophone inventories (Li et al., 2020) and (2) language-agnostic approaches that aim to generalize across languages without such resources (Taguchi et al., 2023; Glocker et al., 2023; Li et al., 2021; Zhu et al., 2025). 🐁POWSM follows the latter paradigm as a fully data-driven multilingual model that learns phone representations without predefined phoneme mappings.

WhisperPPT (Samir et al., 2025) improved Whisper (Radford et al., 2023)'s performance through data cleaning but remained limited in data coverage and task diversity. However, Whisper is trained on a closed corpus and could display harmful biases for PR which cannot be fully removed by fine-tuning. 🐁POWSM is trained from scratch on open datasets.

ZIPA (Zhu et al., 2025) scaled PR to 17,000+ hours of data and 88 languages using a Zipformer (Yao et al., 2024) encoder and noisy-student training on 4,000+ languages, achieving state-of-the-art results. To construct its training corpus, ZIPA employed a G2P system to convert large-scale ASR transcriptions into phoneme sequences, effectively repurposing ASR datasets for PR. Building on this idea, 🐁POWSM leverages both the grapheme and the G2P-generated phoneme transcriptions, reformulating them into four task-specific forms: ASR, PR, G2P, and P2G.

G2P & P2G 🐁POWSM is the first model capable of both audio-guided G2P and audio-guided P2G. G2P conversion, sometimes called phonemization in the text-to-speech literature, can be accomplished with pronunciation dictionaries (Rudnicky, 1993), rules (Mortensen et al., 2018), WFSTs (Black and Lenzo, 2001), or seq2seq neural methods to choose between different pronunciations of a word in context (Zhu et al., 2022). Text-based G2P, however, still cannot handle phonetic

variation, enforcing a one-to-one mapping between orthography and transcription. In contrast, audio-guided G2P can learn to map the different acoustic realizations of a phoneme across varieties of a language to a phone representation (Route et al., 2019). In particular, Mak et al. (2025) observed a performance improvement in using audio-guided G2P versus text-based G2P alone for Cantonese. Gao et al. (2024) similarly showed that joint learning of G2P, phone recognition, and forced alignment outperform a G2P teacher model. Similarly, Sun and Richmond (2024) jointly learned G2P and TTS. Compared to G2P, P2G conversion is less studied, with Lauc (2024) training a seq2seq model on 19 million language-grapheme-phoneme triplets.

3 Methodology

3.1 Data preparation

We use IPAPack++ (Zhu et al., 2025) for training. It is an open source corpus of roughly 17,000 hours of multilingual speech with paired orthographic and phonemic transcriptions. We will release all data processing scripts to make 🐁POWSM fully reproducible.

G2P-generated transcriptions have been manually inspected and cleaned. Following Samir et al. (2025), we remove Interlingua and 10 noisy FLEURS languages. Utterances longer than 300 phones are filtered out. IPA sequences are normalized to Unicode NFD (Canonical Decomposition); English G2P sequences are further refined with rule-based corrections to fix voice-onset time issues (see Appendix § A.1).

To prevent IPA tokens from being confused with graphemes, sequences are split into phone tokens with diacritics and modifiers attached, following PanPhon (Mortensen et al., 2016), and enclosed in slashes (e.g., /pʰɔsəm/ → /pʰ/ /ɔ/ /s/ /ə/ /m/).

3.2 Multitask data format

Our model is trained on four tasks: PR, ASR, and audio-guided G2P and P2G. Each utterance is used once per task, with task-specific formatting as illustrated in Figure 1, including a text prompt, language token, task token, and target output. We leave the text prompt blank (token <na>) for PR and ASR, and provide graphemes and phones as prompts for G2P and P2G.

3.3 Training details

🐁POWSM adopts an attention-based encoder-decoder (AED) architecture, which flexibly models output sequences and allows the integration of additional tasks. Specifically, we follow the OWSM v3.1 architecture (Peng et al., 2024), which employs an E-Branchformer encoder and a Transformer decoder, consistent with the general encoder-decoder structure of Whisper (Radford et al., 2023). The model is trained from scratch using ESPnet (Watanabe et al., 2018) with a hybrid CTC/attention loss (Watanabe et al., 2017), where we set the ratio α_{ctc} to 0.3:

$$\mathcal{L} = \alpha_{\text{ctc}}\mathcal{L}_{\text{ctc}} + (1 - \alpha_{\text{ctc}})\mathcal{L}_{\text{attention}}. \quad (1)$$

The encoder operates at the stride size of 40ms. Training uses a global batch size of 256. Speech inputs are 16kHz and padded to 20 seconds. The vocabulary consists of 40k tokens, including around 6k phone tokens, language and timestamp tokens, and BPE tokens from orthography. The model has approximately 350M parameters with 9 layers for both the encoder and decoder and was trained on 4 H100 GPUs for 2 days. Using a CTC loss (Graves, 2006), We align the encoder outputs with a simplified version of the phone token sequences. Unlike the decoder outputs, the phones in these sequences are stripped of break (/./, /‿/) and length diacritics (/eː/, /eˑ/, /ĕ/) to accelerate convergence. Additional details and analyses are provided in § 6.1.

4 Experimental Setup

Evaluation metric We report Phonetic Feature Error Rate (PFER), an edit distance using articulatory features from PanPhon (Mortensen et al., 2016), averaged over the number of phones and computed as in Equation 2 for PR. Each feature contributes $\frac{1}{24}$ distance unit, while insertion and deletion cost 1 unit. The edit distance D grows linearly with the sequence length and has no upper bound.

$$\text{PFER} = \frac{1}{\#\text{phone}} \sum_{i=1}^{N} D(\text{feat}(\text{hyp}_i), \text{feat}(\text{ref}_i)) \quad (2)$$

Unlike Phone Error Rate (PER), which considers only exact phone matches, or Phone Token Error Rate (PTER), which treats diacritics and modifiers as separate tokens, PFER computes the edit distance in terms of articulatory features—interpretable subphone attributes (*e.g.*

voicing)—capturing phonetic similarity in a fine-grained fashion. Previous studies (Taguchi et al., 2023; Zhu et al., 2025) define PFER as the mean articulatory feature edit distance over the evaluation set. In contrast, we normalize it by the number of phones in the reference transcription to measure the proportion of feature errors per phone.

Decoding hyperparameters We use a CTC weight (denoted as `ctc`) of 0.3 and a beam size (denoted as `beam`) of 3 during decoding for all reported numbers unless specified. Further details on the choice of hyperparameters are discussed in § 6.1.

Evaluation datasets For unseen languages, we evaluate on three datasets: DoReCo (Paschen et al., 2020), VoxAngeles (Chodroff et al., 2024), and Tusom2021 (Mortensen et al., 2021). DoReCo is a dataset of 50+ languages (with broad transcriptions) intended for documentation of small or endangered languages; we use a 45-language subset. VoxAngeles (Chodroff et al., 2024) is a postprocessed version of the UCLA Phonetics Lab Archive (Ladefoged et al., 2009) containing 95 languages. Tusom is a low-data Tangkhulic language of India not included in the training data. Tusom2021 consists of narrow phonetic transcriptions (unlike the broad transcriptions from G2P on which 🎳POWSM was trained) of individual Tusom words.

We also test on five datasets on varieties of English: the Buckeye Corpus (Pitt et al., 2005) and DoReCo South-England represent dialectal variation, while L2-ARCTIC (Zhao et al., 2018), EpaDB (Vidal et al., 2019), and SpeechOcean762 (Zhang et al., 2021) contain L2 speakers. For L2-ARCTIC, we used the manually annotated phoneme transcriptions (which Zhu et al. (2025) termed *L2-Perceived*) rather than G2P dictionary-based transcriptions. The manual transcriptions reflect what the speaker actually said, whereas the dictionary-based version enforces a single pronunciation variant.[1] Manual inspection by a trained phonologist further showed the L2-ARCTIC transcriptions to be of extremely poor quality. For the five aforementioned datasets, we use preprocessed datasets from Zhu et al. (2025)[2] and Koel Labs[3] for better transcription quality.

We then evaluated our model on in-domain data from IPAPack++, the dataset seen during training. We followed Zhu et al. (2025) in using LibriSpeech for English, AISHELL for Mandarin, and MLS for European languages, and additionally evaluated on IISc-MILE Tamil (A et al., 2022) for Tamil and KSC (Khassanov et al., 2021) for Kazakh. For ASR and P2G, we evaluate with FLEURS.

See Table 1 for more details about our evaluation datasets.

PR (In-domain)					
eng	deu	nld	fra	ita	spa
10.58	14.27	12.76	10.07	5.27	10.00
por	pol	tam	kaz	cmn	
3.74	2.14	16.58	7.07	10.02	
PR (Out-of-domain: Unseen languages)					
DoReCo	VoxA.	Tusom.			
19.18	1.58	1.16			
PR (Out-of-domain: Language variation)					
Buckeye	DRC-SE	L2-ARC	EpaDB	SO762	
7.88	0.77	3.66	2.74	2.32	
ASR (FLEURS)					
afr	orm	aze	pan	tgk	mkd
0.66	0.13	2.37	1.48	1.96	2.45
bos	slv				
2.45	1.76				

Table 1: Duration of the test sets for different tasks (in hours). Abbreviated datasets (in order): VoxAngeles, Tusom2021, DoReCo South-England, L2-ARCTIC, SpeechOcean762.

Baselines We evaluate all PR baselines without further training with IPAPack++. See Appendix § A.2 for more details about training data and language coverage. Allosaurus (Li et al., 2020, 2021) uses a phone-level CTC to train a language-agnostic model and applies language-specific allophone-to-phoneme mappings. Wav2Vec2Phoneme (Xu et al., 2022), MultIPA (Taguchi et al., 2023) and Allophant (Glocker et al., 2023) fine-tune XLS-R (Babu et al., 2022) with different objectives: Wav2Vec2Phoneme maps unseen phonemes using articulatory features, MultIPA leverages high-quality G2P data from seven languages, while Allophant decomposes phones into articulatory features and applies CTC losses for each. ZIPA (Zhu et al., 2025) trains ZipFormer

[1] For instance, "crayon" in American English can be pronounced as /ˈkɹæn/, /ˈkɹej.ɔn/, or /ˈkɹej.ɒn/ (Vaux and Golder, 2003) (among others), but the CMU Pronouncing Dictionary (Rudnicky, 1993) only lists one.

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

[3] https://huggingface.co/KoelLabs

(Yao et al., 2024) from scratch on IPAPack++ using CR-CTC and also provides a variant trained with additional pseudo-labeled data ("ZIPA-CR-NS-Large").

For ASR, we compare 🐁POWSM with two series of models: OWSM (Peng et al., 2025) and OWLS (Chen et al., 2025). We select OWSM-CTC v4 because it is the best-performing model in the series, featuring an encoder-CTC architecture that supports ASR, ST, and LID. For OWLS, we include models with comparable parameter sizes.

5 Results

We found that 🐁POWSM 's performance on PR and ASR tasks is comparable or superior to competitive baselines.

5.1 Multi-task performance

Results on the in-domain test sets are presented in Table 2 and Table 3. We provide further discussion of G2P and P2G in § 6.2.

🐁POWSM excels at in-domain phone recognition From Table 2, we see that 🐁POWSM achieves the lowest average PFER in phone recognition, due to the strong language modeling capability of the decoder. We hypothesize that our English data cleaning (Appendix § A.1) may have negatively affected the PFER for Germanic languages due to a mismatch between training and test data. Nevertheless, our approach fills this gap by achieving strong performance on other languages, outperforming models trained on larger datasets.

🐁POWSM is comparable with web-scale ASR models on low-resource languages We hypothesize that pre-training with phone recognition benefits low-resource ASR (Yusuyin et al., 2025). To choose low-resource languages, we selected languages in IPAPack++ with less than 8 hours of speech in FLEURS to serve as the test set. See § A.3 for details on the amount of data used by different models. For a fair comparison with other multilingual ASR baselines without language-specific components, we use the same decoding hyperparameters `ctc=0.0, beam=1`.

As shown in Table 3, 🐁POWSM (`POWSM 0.35B, ASR`) is often comparable to models of similar size trained on web-scale data for ASR (`OWLS 0.5B`). Incorporating phones obtained from PR as text prompts (PR-P2G) significantly decreases WER, making it comparable to or even better than these models. When using gold phone labels for P2G (see analysis in § 6.2), 🐁POWSM outperforms other ASR models by a large margin in most cases.

5.2 🐁POWSM generalizes well to unseen languages

Table 4 reports PFER on datasets with unseen languages and language variation. Results indicate that 🐁POWSM achieves strong performance across these datasets, and handles both dialectal and L2 variation effectively. Notably, our method outperforms ZIPA trained on the same data and even exceeds ZIPA trained with extra pseudo-labeled data, achieving the best results on unseen languages while performing three additional tasks. This shows the effectiveness of our multi-task approach. While 🐁POWSM lags behind Wav2Vec2Phoneme on socio-phonetic variations, we attribute this to its self-supervised learning with over 60k hours of speech (from wav2vec 2.0 (Baevski et al., 2020)) prior to the supervised learning stage.

6 Analysis

In this section, we analyze how 🐁POWSM works, focusing on the phonetic-aware encoder and task- and language-specific tokens, which are the defining features of the model.

6.1 Behavior of the speech encoder

The CTC encoder prefers fine-grained phones without suprasegmentals We observed that mixing phones and orthography as encoder targets hindered training, because the same speech input would have different encoder CTC targets for different tasks. Therefore, we used phones as encoder targets, encouraging general representations of sounds to be shared across languages.

To determine the most effective unit for the CTC encoder, we fix the decoder vocabulary to PanPhon phones and compared four encoder targets: (1) Unicode code points vs. PanPhon, and (2) sequences with vs. without suprasegmentals (length and break marks). Unicode code points offer simplicity and a smaller vocabulary but split phones into unnatural units (e.g. /p//ʰ/) and increase sequence length, while PanPhon represents each phone-diacritic combination as a unit (e.g. /pʰ/), yielding a more natural monotonic sequence at the expense of sparsity and potential out-of-vocabulary

Model	Param.	eng	deu	nld	fra	ita	spa	por	pol	tam	kaz	cmn	Avg.
Allosaurus	11M	6.89	17.67	19.19	20.91	19.02	4.82	19.61	21.21	12.01	20.90	15.28	16.14
Allophant	300M	10.26	9.37	18.39	18.83	7.82	17.37	15.44	7.90	19.32	—	—	—
Wav2Vec2Phoneme	300M	7.70	7.89	12.31	17.73	6.10	3.67	11.65	9.57	15.63	15.30	14.66	11.11
MultIPA	300M	15.81	16.28	18.97	20.19	7.20	6.99	15.04	2.63	10.54	17.71	21.10	13.86
ZIPA-CR-Large	300M	1.63	3.32	3.03	3.23	3.24	1.98	4.01	4.33	4.59	2.31	1.25	2.99
ZIPA-CR-NS-Large	300M	**1.40**	**3.17**	**2.83**	**2.92**	3.22	1.53	3.40	4.31	4.15	**1.87**	**0.90**	2.70
POWSM	350M	2.85	3.37	5.14	3.27	**1.81**	**1.21**	**2.90**	**1.36**	**3.56**	2.25	1.15	**2.62**

Table 2: PFER (↓) on the in-domain dataset, IPAPack++. Languages not supported by Allophant are left blank. Some languages were not seen by MultiIPA. **Bold** indicates the best performance.

	Afroasiatic		Turkic	Indo-Iranian		Balto-Slavic		
Model	afr	orm	aze	pan	tgk	mkd	bos	slv
OWLS 0.5B	102.3	**89.0**	77.7	59.3	60.4	54.2	59.3	58.6
OWLS 1B	95.7	102.4	67.5	**50.0**	**50.7**	**46.2**	**50.0**	**52.8**
OWSM-CTC v4 1B	**67.5**	92.7	71.2	88.7	57.6	51.2	51.3	60.4
POWSM 0.35B, ASR	86.2	125.3	67.7	83.1	62.8	56.0	56.5	64.5
POWSM 0.35B, PR-P2G	68.8	93.0	**66.7**	72.8	51.0	48.6	56.9	63.9

Table 3: WER (↓) of ASR and PR-P2G on low-resource languages. PR-P2G uses phones predicted by PR as text prompts instead of gold phones. **Bold** indicates the best performance, and underline indicates the second-best.

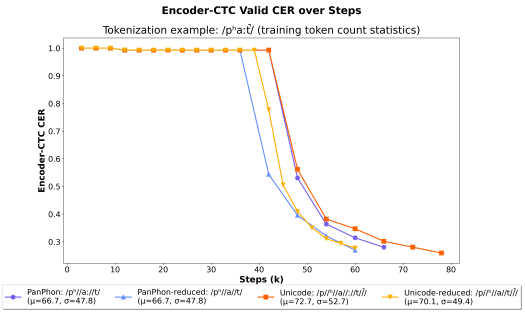

Figure 2: Validation CER of encoder-CTC during training. Removing suprasegmentals for CTC accelerates convergence and reduces vocabulary size for encoder.

issues. Suprasegmentals such as /ː/, though phonemic in many languages, confuse PR models (Zhu et al., 2025).

We run small-scale experiments on a 1k-hour subset of the multi-task data (250 hours of speech repeated across four tasks). We use the validation CER of the encoder-CTC output as a proxy for training efficiency. An earlier drop indicates that the encoder is learning a useful alignment early, which improves representations fed into the decoder and accelerates overall convergence. In Figure 2, PanPhon tokenization without suprasegmentals shows the earliest drop, suggesting that alignment with decoder units aids training, while collapsing suprasegmental distinctions for CTC reduces confusion.

Increased encoder weights benefit PR on out-of-domain data As in other encoder-decoder models (Gong et al., 2023; Radford et al., 2023), we expect the encoder of 🐁 POWSM to capture more general acoustic patterns, while the decoder handles language and task-specific output formats. Therefore, we investigate whether emphasizing the encoder more during different stages of model development affects performance. To balance data diversity with inference compute costs, we selected two smaller datasets from each category with distinct characteristics.

As shown in Table 5, higher CTC decoding weights improve PR performance on out-of-domain data but degrade it on in-domain data, as expected. This echoes Zhu et al. (2025)'s finding that RNN-T (Graves and Jaitly, 2014), an encoder-only speech-to-text model with an autoregressive text prediction network, hurts generalization to unseen patterns of phones (phonotactics). We hypothesize that the decoder is performing implicit language modeling and "smooths" phonetic variation, as Zhu et al. (2025) described.

Next, we examine whether focusing more on the CTC loss through training widens this gap in performance between in-domain and out-of-domain data. We find that fine-tuning with a higher CTC loss weight α_{ctc} after convergence does not improve out-of-domain performance and can even degrade it. Randomly varying α_{ctc} for each batch also shows no improvement. In contrast, training with a higher α_{ctc} from the start benefits the out-of-domain distribution, achieving the lowest PFER on unseen languages with greedy decoding, while the PFER on in-domain data is comparatively higher. These results suggest that assigning a higher weight to the encoder during training and inference improves PR, highlighting a common trade-off between in-domain performance and generalization.

Model	Param.	Unseen Languages				Language Variation					
		DoReCo	VoxAngeles	Tusom2021	Avg.	Buckeye	DRC-SE	L2-ARC	EpaDB	SO762	Avg.
Allosaurus	11M	24.71	30.84	42.02	32.52	15.24	25.36	13.39	19.33	21.61	18.99
Allophant	300M	—	—	—	—	16.05	24.13	11.91	14.38	18.28	16.95
Wav2Vec2Phoneme	300M	17.25	**13.88**	31.92	21.02	12.50	18.57	9.86	**9.90**	**13.60**	**12.89**
MultIPA	300M	18.28	15.23	30.53	21.35	18.69	23.31	15.52	15.64	21.34	18.90
ZIPA-CR-Large	300M	17.99	16.95	23.68	19.54	**12.04**	17.89	9.74	17.38	15.58	14.53
ZIPA-CR-NS-Large	300M	16.82	17.14	23.08	19.01	12.05	**17.12**	**9.69**	14.63	18.20	14.34
POWSM	350M	**17.06**	17.11	**21.96**	**18.71**	12.63	18.33	11.32	11.86	17.84	14.40

Table 4: PFER (↓) on out-of-domain data. "DRC-SE" stands for DoReCo South-England; "L2-ARC" stands for L2-ARCTIC; "SO762" stands for SpeechOcean762. Unseen language datasets include languages not supported by Allophant; therefore, we do not report results for these datasets.

Setup	In-domain		Out-of-domain			
	ita	pol	VoxA.	Tusom.	DRC-SE	EpaDB
Decoding						
ctc=0.3	**1.66**	**1.36**	**17.58**	33.52	18.21	11.88
ctc=0.7	1.97	1.37	17.92	24.29	18.05	11.82
ctc=0.9	2.05	1.38	19.27	**22.94**	**17.67**	**11.80**
Pre-training / Fine-tuning						
α_{ctc}=0.3	**1.81**	1.60	16.02	22.47	18.59	11.66
→ Ft, α_{ctc}=0.5	1.94	**1.53**	17.78	23.72	18.73	11.82
α_{ctc}=0.7	1.96	1.65	**15.40**	**22.10**	18.50	11.62
→ Ft, α_{ctc}=0.5	2.01	1.57	16.21	22.93	**18.41**	11.47
$\alpha_{\text{ctc}}=U(0.1,0.9)$	1.95	1.62	19.29	25.11	18.92	**11.33**

Table 5: PFER (↓) for different CTC weight settings. "Ft" denotes fine-tuning for 5 epochs from the checkpoint above. VoxAngeles and Tusom2021 are abbreviated. Pre-training and fine-tuning rows use `ctc=0.3`. All setups use `beam=1`.

Task	Buckeye	Example
ASR Transcription		any holidays at all they just kind of ignore
Phonetic transcription		/ɛnihɑlʌdeɪsɛɾɑlsoʊðɛɪdʒʌstkɑɾʌvɪgnɔɹ/
PR	12.63	/ɛ̃nihɑlədeɪzætəlsoʊðɛɪtʃʌstkʰɑ̃ɛ̃nəvɪgnɔɹ/
G2P (speech)	12.71	/ɛ̃nihɑlədeɪzætəlsoʊðɛɪtʃʌstkʰɑ̃ɛ̃nəvɪgnɔɹ/
G2P (both)	16.38	/ɛ̃nihɑlədeɪzætəlðɛɪtʃʌstkʰĩndəvɪgnɔɹ/
G2P (text prompt)	23.44	/aɪhoʊldaɪzætəlðɛɪtʃɪstkʰĩndəvɪgnɜ/

Table 6: Comparing G2P with different available modalities with PFER (↓). Blue for correctly capturing mispronounced parts (/soʊ/), orange for error compared to other examples.

6.2 Inspecting Task and Language Tokens

Speech-guided G2P preserves phonetic variation; text prompts normalize it To better understand how 🐁 POWSM integrates speech and text prompts, we analyze the relative influence of speech and text prompts in its G2P behavior. We vary the G2P conditions from purely speech-based to purely text-based, as shown in Table 6, and evaluate the model on the Buckeye dataset. When only speech is provided, the performance is comparable to the PR setting, which differs only in the task token. Adding both speech and text prompts (the standard G2P setup) leads to degraded performance, with output showing standardized pronunciations. When the model relies solely on the text prompt, performance drops sharply and pronunciations become highly standardized as expected (just as Zhu et al. (2025) reported).

In other words, 🐁 POWSM G2P responds to speech and text signals to controllably mediate between narrow and broad transcription. In the multi-task setup, this effect may be stronger because the model is trained with G2P, which could bias it toward more standardized forms.

Audio-P2G effectively handles low-resource languages We compare several P2G setups on the same set of low-resource languages from FLEURS, listed in Table 7. P2G significantly outperforms ASR, suggesting that it effectively leverages the provided phone context. However, since P2G uses gold phone labels, this comparison is not entirely fair. We therefore tested PR followed by P2G (PR-P2G), and found that performance improved for some languages but not for others. Error propagation does not explain this variation in performance, as PFER trends from PR differ from the observed performance drops. Yet the PFER pattern aligns closely with ASR results, suggesting that phonotactic similarity to high-resource languages may play a role.

To test this, we run P2G with the language code set to English and post-process the output to match Cyrillic or Gurmukhi transcriptions with online conversion tools for certain languages.[4] This approach often outperforms ASR and sometimes approaches P2G's performance, indicating that P2G also relies heavily on speech input. Languages with either comparibly low or high PFER did not benefit from this transliteration approach, possibly because the model already handled them well or had not yet learned them sufficiently. This finding suggests a

[4] https://www.lexilogos.com for Macedonian and Tajik; https://punjabi.indiatyping.com for Panjabi.

	Afroasiatic		Turkic	Indo-Iranian		Balto-Slavic		
Task	afr	orm	aze	pan	tgk	mkd	bos	slv
Best ASR	67.5	89.0	67.5	**50.0**	50.7	46.2	50.0	52.8
ASR	86.2	125.3	67.7	83.1	62.8	56	56.5	64.5
P2G	**55.9**	**88.0**	**37.1**	52.6	**31.8**	**36.9**	**32.3**	**40.3**
P2G, lang=<eng>	60.4	99.0	64.2	*95.8	*74.0	*52.6	39.6	53.5
PR-P2G	68.8	93.0	66.7	72.8	51.0	48.6	56.9	63.9
PFER (↓)	9.1	12.9	6.7	7.9	5.7	3.3	6.7	6.9

Table 7: WER (↓) of different P2G settings on low-resource languages "Best" stands for lowest WER in Table 3 from ASR models. * indicates post-processed languages.

direction for further investigation in low-resource ASR.

Language token captures phonotactics The language identification (LID) performance of 🐁POWSM on seen languages in FLEURS reaches 92.3% accuracy, as shown in Figure 3. To see if the model implicitly learns phonotactic patterns and associates them with the language token, we evaluate PR on unseen languages by manipulating the language token at inference time. For VoxAngeles and Tusom2021, the three most frequently assigned languages are Bashkir (42.6%, 25.1%), English (30.2%, 67.7%), and Kinyarwanda (14.5%, 2.3%), which are all relatively high-resource languages in IPAPack++. Table 8 shows that assigning the detected language token yields better performance than always using English, while setting the language as unknown performs best. This indicates that the language token influences PR by shifting the output distribution toward the assigned language.

Lang. token	Voxangeles	Tusom2021	Example
Phonetic transcription			/adʒmə/
<unk>	17.11	21.96	/ad͡ʒima/
Detected	17.55	23.92	/ɒjum/
<eng>	19.91	24.21	/aɪtimɔ/

Table 8: PFER (↓) of PR performance with different language token. The detected language in the example is <bak>. Blue for correct, orange for error compared to other examples.

7 Conclusion

We train a fully open-source phonetic speech foundation model 🐁POWSM using our scalable multi-task framework. Our model achieves state-of-the-art performance on PR while also supporting ASR across more than 70 languages. Beyond PR and ASR, the model's ability to perform audio-guided G2P and P2G enables applications that require fine-grained linguistic analysis such as atypical speech assessment. Our analysis reveals that 🐁POWSM 's encoder benefits from phoneme-level CTC supervision and stronger encoder weighting, enhancing cross-lingual generalization. Additionally, the model demonstrates interpretable multimodal and language-aware behaviors, effectively mediating between phonetic detail and standardized phonological patterns. To conclude, 🐁POWSM not only provides a strong phone recognition foundation model for high-resource languages, but also acts as a versatile resource for unseen languages and socio-phonetic variation.

8 Future Work

🐁POWSM 's current decoder serves as a large phoneme-level phonotactic language model on which linguists could investigate hypotheses about phonetic universals (Chodroff et al., 2024; Chodroff, 2025) and phonotactics (Shim et al., 2024; Pimentel et al., 2020). In the future, we seek to adapt to socio-phonetic variation either through (unsupervised) test-time adaptation (Lin et al., 2022), in-context learning (Roll et al., 2025; Wang et al., 2024), or mechanistic interpretability (Tang et al., 2024). Furthermore, since Shim et al. (2025) found that earlier encoder layers in Whisper preserve more phonetic detail, early exiting may mitigate the decoder's tendencies to normalize socio-phonetic variation.

Limitations

🐁POWSM has several limitations that we aim to address in future work. First, the model is neither strictly phonemic nor phonetic: its training data consist of cleaned and filtered phonemic transcriptions from multiple languages, which are not fully faithful to the phonetic or phonemic structure of the audio. Although phonemic transcriptions share similarities across languages, adding auxiliary tasks and language tokens may have reinforced language-specific biases. We also currently lack sufficient allophone-level data, which would provide more language-independent information.

Second, the model still favors high-resource languages. Since we include a decoder for language modeling and language tokens, both of which function effectively, the model would inherently bias toward the seen distribution.

Finally, the current AED architecture imposes engineering limitations. Inference is significantly slower than with encoder-only models, and the architecture does not easily support tone modeling, limiting its application to tonal languages.

Ethics Statement

All of our data is ethically sourced, either through permissive licensing or through proper consent. We are aware of the implicit prescriptivism and representational harms (Crawford, 2017) that normalizing socio-phonetic variation in ASR or PR models can create. This may threaten linguistic diversity instead of preserving it. We also acknowledge that accurate modeling of socio-phonetic variation can enable demographic inference, as demographics and phonetic variation are deeply intertwined (Labov, 1963). We stress that uses of 🐁POWSM must align with our vision: a future where advances in spoken language processing and NLP do not leave low-resource varieties behind.

The Use of LLMs

We acknowledge the use of large language models (LLMs) to assist with grammar correction and clarity improvements in writing this paper. All conceptual, methodological, and experimental contributions were developed independently by the authors.

References

Madhavaraj A, Bharathi Pilar, and Ramakrishnan A G. 2022. Subword dictionary learning and segmentation techniques for automatic speech recognition in tamil and kannada. *Preprint*, arXiv:2207.13331.

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. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Interspeech 2022*, pages 2278–2282.

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.

Yonatan Belinkov and James Glass. 2017. Analyzing hidden representations in end-to-end automatic speech recognition systems. *Advances in Neural Information Processing Systems*, 30.

Alan W Black and Kevin A Lenzo. 2001. Flite: a small fast run-time synthesis engine. In *SSW*, page 204.

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

Eleanor Chodroff. 2025. Phonetic universals. *Annual Review of Linguistics*, 11.

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, 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.

Kate Crawford. 2017. The trouble with bias. Conference on Neural Information Processing Systems.

Li Fu, Yu Xin, Sunlu Zeng, Lu Fan, Youzheng Wu, and Xiaodong He. 2025. Pac: Pronunciation-aware contextualized large language model-based automatic speech recognition. *arXiv preprint arXiv:2509.12647*.

Heting Gao, Mark Hasegawa-Johnson, and Chang D Yoo. 2024. G2pu: Grapheme-to-phoneme transducer with speech units. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10061–10065. IEEE.

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.

Kevin Glocker, Aaricia Herygers, and Munir Georges. 2023. Allophant: Cross-lingual phoneme recognition with articulatory attributes. In *Interspeech 2023*, pages 2258–2262.

Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass. 2023. Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers. In *Interspeech 2023*, pages 2798–2802.

A Graves. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. Int. Conf. on Machine Learning, 2006*, pages 369–376.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772. PMLR.

Gerard I. Gállego, Oriol Pareras, Martí Cortada Garcia, Lucas Takanori, and Javier Hernando. 2025. Speech-to-text translation with phoneme-augmented cot: Enhancing cross-lingual transfer in low-resource scenarios. *Preprint*, arXiv:2505.24691.

Taiqi He, Kwanghee Choi, Lindia Tjuatja, Nathaniel Robinson, Jiatong Shi, Shinji Watanabe, Graham Neubig, David Mortensen, and Lori Levin. 2024. Wav2Gloss: Generating interlinear glossed text from speech. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 568–582, Bangkok, Thailand. Association for Computational Linguistics.

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.

Yerbolat Khassanov, Saida Mussakhojayeva, Almas Mirzakhmetov, Alen Adiyev, Mukhamet Nurpeiissov, and Huseyin Atakan Varol. 2021. A crowdsourced open-source Kazakh speech corpus and initial speech recognition baseline. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 697–706, Online. Association for Computational Linguistics.

William Labov. 1963. The social motivation of a sound change. *Word*, 19(3):273–309.

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.

Davor Lauc. 2024. Polyipa–multilingual phoneme-to-grapheme conversion model. *arXiv preprint arXiv:2412.09102*.

Chin-Jou Li, Eunjung Yeo, Kwanghee Choi, Paula Andrea Pérez-Toro, Masao Someki, Rohan Kumar Das, Zhengjun Yue, Juan Rafael Orozco-Arroyave, Elmar Nöth, and David R Mortensen. 2025. Towards inclusive asr: Investigating voice conversion for dysarthric speech recognition in low-resource languages. *arXiv preprint arXiv:2505.14874*.

Kun Li, Xiaojun Qian, and Helen Meng. 2016. Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1):193–207.

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, Juncheng Li, Florian Metze, and Alan W Black. 2021. Hierarchical phone recognition with compositional phonetics. In *Interspeech*, pages 2461–2465.

Guan-Ting Lin, Shang-Wen Li, and Hung yi Lee. 2022. Listen, Adapt, Better WER: Source-free Single-utterance Test-time Adaptation for Automatic Speech Recognition. In *Interspeech 2022*, pages 2198–2202.

Ryo Magoshi, Shinsuke Sakai, Jaeyoung Lee, and Tatsuya Kawahara. 2025. Multi-lingual and Zero-Shot Speech Recognition by Incorporating Classification of Language-Independent Articulatory Features. In *Interspeech 2025*, pages 91–95.

Timothy Shin Heng Mak, King Yiu Suen, and Albert Lam. 2025. Speech-guided grapheme-to-phoneme conversion for cantonese text-to-speech. In *Proc. Interspeech 2025*, pages 2535–2539.

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.

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.

Yifan Peng, Muhammad Shakeel, Yui Sudo, William Chen, Jinchuan Tian, Chyi-Jiunn Lin, and Shinji Watanabe. 2025. OWSM v4: Improving Open Whisper-Style Speech Models via Data Scaling and Cleaning. In *Interspeech 2025*, pages 2225–2229.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel,

Kwanghee Choi, Jiatong Shi, Xuankai Chang, Jee weon Jung, and Shinji Watanabe. 2024. OWSM v3.1: Better and Faster Open Whisper-Style Speech Models based on E-Branchformer. In *Interspeech 2024*, pages 352–356.

Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, and 1 others. 2023. Reproducing whisper-style training using an open-source toolkit and publicly available data. In *2023 IEEE ASRU*, pages 1–8. IEEE.

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

Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. 2005. The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

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

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.

Nathan Roll, Calbert Graham, Yuka Tatsumi, Kim Tien Nguyen, Meghan Sumner, and Dan Jurafsky. 2025. In-context learning boosts speech recognition via human-like adaptation to speakers and language varieties. *arXiv preprint arXiv:2505.14887*.

James Route, Steven Hillis, Isak Czeresnia Etinger, Han Zhang, and Alan W Black. 2019. Multimodal, multilingual grapheme-to-phoneme conversion for low-resource languages. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 192–201, Hong Kong, China. Association for Computational Linguistics.

Alexander Rudnicky. 1993. The cmu pronouncing dictionary. Accessed on October 2, 2025.

Farhan Samir, Emily P. Ahn, Shreya Prakash, Márton Soskuthy, Vered Shwartz, and Jian Zhu. 2025. A comparative approach for auditing multilingual phonetic transcript archives. *Transactions of the Association for Computational Linguistics*, 13:595–612.

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.

Ryan Soh-Eun Shim, Domenico De Cristofaro, Chengzhi Martin Hu, Alessandro Vietti, and Barbara Plank. 2025. Languages in multilingual speech foundation models align both phonetically and semantically. *arXiv preprint arXiv:2505.19606*.

Siqi Sun and Korin Richmond. 2024. Acquiring pronunciation knowledge from transcribed speech audio via multi-task learning. *arXiv preprint arXiv:2409.09891*.

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.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438*.

Bert Vaux and Scott Golder. 2003. The Harvard Dialect Survey.

Jazmín Vidal, Luciana Ferrer, and Leonardo Brambilla. 2019. Epadb: A database for development of pronunciation assessment systems. In *INTERSPEECH*, pages 589–593.

Siyin Wang, Chao-Han Yang, Ji Wu, and Chao Zhang. 2024. Can whisper perform speech-based in-context learning? In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13421–13425. IEEE.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, and 1 others. 2018. Espnet: End-to-end speech processing toolkit. *Interspeech 2018*.

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.

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

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*.

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*.

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

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

Xuehao Zhou, Xiaohai Tian, Grandee Lee, Rohan Kumar Das, and Haizhou Li. 2020. End-to-end code-switching tts with cross-lingual language model. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7614–7618.

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

A Appendix

A.1 Refining English G2P

We observed confusion in plosive voice-onset times on unseen languages in preliminary experiments, which is likely from English G2P data. For instance, broad phonemic transcription in English typically uses /b/ to transcribe the /b/ in /bat/, but its voice onset timing is actually voiceless in Mainstream American English and is closer to [p]. To mitigate this, we apply rule-based refinements to English G2P transcriptions, adjusting plosive voicing and aspiration, lateral velarization, and vowel nasalization.

The rules are listed below: 1) word-initial voiceless plosives (/p/, /t/, /k/) are aspirated, 2) word-initial voiced plosives (/b/, /d/, /g/) are voiceless, 3) lateral /l/ is velarized at the end of syllables, and 4) vowel nasalization before nasal consonants.

A.2 Baseline Implementation

We provide the baselines' training data source, number of languages covered in the data, and links to model checkpoints or repository in Table 9.

A.3 FLEURS language selection for ASR

We first filter out languages with more than 8 hours of training data in IPAPack++ (Zhu et al., 2025), keeping only those that are also present in FLEURS. Then, following the training data amounts reported in Chen et al. (2025), we further identify the 50 lowest-resource languages to exclude any that may have other substantial sources not included in IPAPack++. This process leaves us with nine languages. We finally exclude `ell`, as it is comparatively higher-resource and because there are already three other Balto-Slavic languages. Note that other models use strictly more data than ours—not only in terms of dataset count but also because IPAPack++ applies additional data-quality filtering. Table 10 lists the amount of ASR training data for baselines.

A.4 Multi-tasking at Different Scales

Multi-tasking may improve performance by tying acoustic signals to well-defined symbolic representations, yet it may distract the model if the relationships are not learned effectively. We train 🐁 POWSM with different data and model scales to examine how multitask learning interacts with the setup, and use `beam=1` during decoding to speed up inference.

Table 11 shows that there is no clear trend regarding whether multitasking benefits PR performance. PR performance degrades when the model has excessive capacity relative to the available data (too little data), or when it is limited by size (too much data).

Further evidence is needed before concluding that phoneme recognition benefits less from scaling, as we currently lack sufficient data and large model capacity to test this thoroughly. Nevertheless, the model demonstrates the ability to multitask, which represents a promising direction for future work.

Model	Training Data Sources	Language Coverage	Model checkpoint / GitHub
PR baselines			
Allosaurus (Li et al., 2020, 2021)	VoxForge, Japances CSJ, Hkust Tedlium, Switchboard etc	12	xinjli/allosaurus
Allophant (Glocker et al., 2023)	Common Voice 10.0	34	kgnlp/allophant
Wav2Vec2Phoneme (Xu et al., 2022)	MLS, Common Voice, Babel	40+	facebook/wav2vec2-xlsr-53-espeak-cv-ft
MultIPA (Taguchi et al., 2023)	Common Voice 11.0	7	ctaguchi/wav2vec2-large-xlsr-japlmthufielta-ipa1000-ns
ZIPA (Zhu et al., 2025)	IPAPack++ MMS ulab v2., VoxLingua-107 (Pseudo-label)	88	lingjzhu/zipa anyspeech/zipa-large-crctc-ns-800k
ASR baselines			
OWSM-CTC v4 (Peng et al., 2025)	OWSM v3.2, YODAS	100+	espnet/owsm_ctc_v4_1B
OWLS (Chen et al., 2025)	OWSM v3.2, YODAS	150	espnet/owls-scaling-laws-for-speech-recognition-and-translation

Table 9: Overview of the baselines for our work.

	Afroasiatic		Turkic	Indo-Iranian		Balto-Slavic		
Model	afr	orm	aze	pan	tgk	mkd	bos	slv
POWSM	2.71	5.11	6.89	4.96	6.52	5.14	7.57	7.19
OWSM-CTC v4 OWLS	5.54	6.50	10.69	8.30	8.03	8.4	9.96	26.00

Table 10: Amount of ASR training data for languages included in ASR comparison (in hours), according to Zhu et al. (2025) and Chen et al. (2025).

Data (khr)	Tasks	Params.	VoxAngeles	Tusom2021	L2-Arctic
0.25	1	100M	27.22	32.59	13.50
0.25	4	100M	26.30	30.32	13.40
0.25	1	300M	20.63	25.83	12.88
0.25	4	300M	23.81	25.91	14.14
17	1	100M	17.88	26.68	11.76
17	2	100M	24.69	49.28	11.35
17	4	100M	30.07	61.89	12.37
17	1	300M	17.08	25.20	10.50
17	2	300M	17.17	23.70	10.47
17	4	300M	17.58	33.52	10.54

Table 11: Comparison of PFER (↓) on different seting of tasks and data. 1 task refers to PR, 2 tasks refer to PR+ASR, and 4 tasks include PR, ASR, P2G, and G2P.

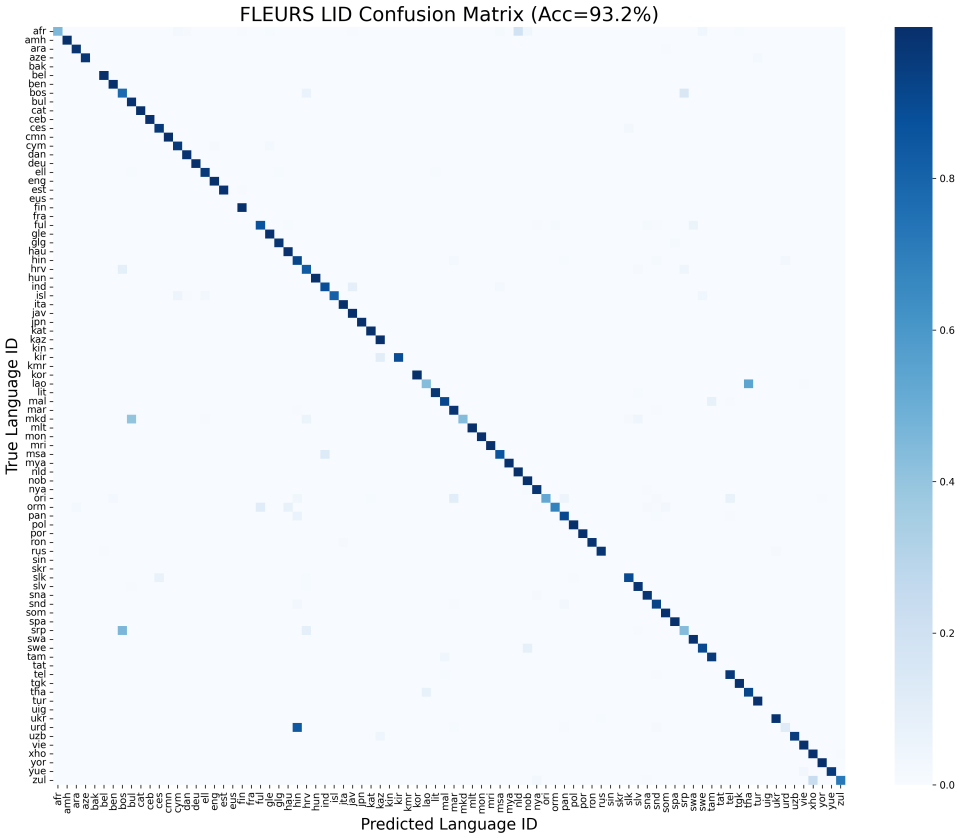

Figure 3: Confusion matrix of LID on FLEURS.