ZIPA: A family of efficient models for multilingual phone recognition

Jian Zhu
University of British Columbia
jian.zhu@ubc.ca

Farhan Samir
University of British Columbia
fsamir@mail.ubc.ca

Eleanor Chodroff
University of Zurich
eleanor.chodroff@uzh.ch

David R. Mortensen
Carnegie Mellon University
dmortens@cs.cmu.edu

Abstract

We present ZIPA, a family of efficient speech models that advances the state-of-the-art performance of crosslinguistic phone recognition. We first curated IPAPACK++, a large-scale multilingual speech corpus with 17,132 hours of normalized phone transcriptions and a novel evaluation set capturing unseen languages and sociophonetic variation. With the large-scale training data, ZIPA, including transducer (ZIPA-T) and CTC-based (ZIPA-CR) variants, leverage the efficient Zipformer backbones and outperform existing phone recognition systems with much fewer parameters. Further scaling via noisy student training on 11,000 hours of pseudo-labeled multilingual data yields further improvement. While ZIPA achieves strong performance on benchmarks, error analysis reveals persistent limitations in modeling sociophonetic diversity, underscoring challenges for future research.

1 Introduction

The International Phonetic Alphabet (IPA) provides a theoretically unified discrete representation of all known human speech sounds (International Phonetic Association, 1999). IPA transcriptions capture major articulatory contrasts in speech sounds, including the voicing status, place of articulation, manner of articulation, and tongue positions (Ladefoged and Johnson, 2014). In phonetics, the IPA is the major tool to document speech sounds across the world's languages, thanks to its universality. Therefore, developing speech technology that can transcribe multilingual speech into phones, or IPA symbols can significantly facilitate language documentation, especially for low-resource languages.

Even beyond linguistics, phone transcriptions are also widely used in various speech technologies, including multilingual pretraining (e.g., Feng et al., 2023; Yusuyin et al., 2025), speech synthesis (e.g., Liu et al., 2023), speech enhancement (e.g., Liu et al., 2021; Pirklbauer et al., 2023), pronunciation assessments (e.g., Zhang et al., 2021; Gong et al., 2022), and voice conversion (e.g., Lee et al., 2022; Shan et al., 2024).

In this study, we present state-of-the-art phone recognition systems that can transcribe speech into IPA symbols crosslinguistically. Our core contributions are summarized as follows.

- First, we curate IPAPACK++, a 17,132-hour open-source speech corpora with G2P-generated phonetic transcriptions. We also design an evaluation set containing rich crosslinguistic and sociophonetic variation.
- Second, we present a series of state-of-the-art phone recognition models, the transducer ZIPA-T and the CTC-based ZIPA-CR in two sizes (64M and 300M). Trained on the IPAPACK++, even the 64M ZIPA models outperform previous phone recognition models with 300M parameters, while being more computationally efficient.
- Third, we further applied noisy student training on ZIPA-CR models with 11k hours of pseudo-labeled speech in more than 4,000 languages, resulting in state-of-the-art performance on phone recognition.
- Finally, we conducted error analyses on the model prediction, showing that current phone recognition models, despite the impressive performance, are still struggling with predicting sociophonetic variation. Our analysis thus reveals a critical, overlooked limitation of current data curation practices in training universal phone recognition models.

We will release all training and evaluation data, pretrained models, and the code under permissive licenses at https://github.com/lingjzhu/zipa.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19568–19585
July 27 - August 1, 2025

2 Background

2.1 Multilingual phone recognition

Early efforts in automatic speech recognition in the 1970s were centered on prediction of phones (Li, 2017). There has been a resurgence in interest in phonetic transcription (Li et al., 2020; Gao et al., 2021; Xu et al., 2022; Taguchi et al., 2023; Glocker et al., 2023; Samir et al., 2024). These models have proven indispensable for transcribing speech in oral languages (Lane and Bird, 2021), and have high potential for facilitating cross-linguistic phonetic analysis (Chodroff et al., 2024). Most systems are trained through fine-tuning pretrained multilingual models like XLS-R and Whisper (Babu et al., 2021; Radford et al., 2023) on large audio-transcript archives like VoxClamantis (e.g., Salesky et al., 2020) or X-IPAPack (Samir et al., 2024). But the transcripts are semi-automatically generated through applying G2P models to orthographic transcripts.

Still, there remain significant challenges with training reliable phonetic transcript models for the world's languages. First, the linguistic diversity of the datasets needs to be considerable in order to transcribe audio from any language. As shown in Samir et al. (2024), collecting reliably transcribed audio-transcript pairs is far from trivial, as algorithmic curation pipelines for obtaining massively multilingual transcribed audio archives can fail. Importantly, these failures manifest when the G2P model is not calibrated for the language variety represented by the audio. To this end, we collect the IPAPACK++ dataset (Section 3), comprising 17K+ hours of reliable phonetically transcribed audio in 88 languages.

Moreover, another potential challenge is that G2P models tend to capture dictionary-like pronunciations for the standard dialect of the language, thereby failing to capture pronunciation patterns in audio for different sociolects. Therefore, we specifically design evaluation datasets rich in sociophonetic variation to evaluate whether the phone recognition models are simply memorizing the standard pronunciations.

2.2 Phone recognition is subjective

While the IPA provides a universal representation of speech sounds, applying IPA crosslinguistically still poses many challenges. The acoustic-phonetic details of a given speech segment can vary considerably across speakers and languages. For example, voice onset time (VOT) is commonly known as the primary acoustic correlate for separating voiceless from voiced stops across languages (Abramson and Whalen, 2017). However, the absolute values of VOT vary substantially across languages (Cho and Ladefoged, 1999; Chodroff et al., 2019), which cannot be easily captured via discrete IPA symbols.

Therefore, phonetic transcription remains a highly subjective process, affected by the linguistic backgrounds or theoretical orientations of the transcriber. In transcription practices, strict transcriptions are not always necessary or achievable because many non-contrastive phonetic details are usually irrelevant in a given analysis linguistic analysis (Anderson et al., 2023; Kerswill and Wright, 1990; Shriberg and Lof, 1991). Shriberg and Lof (1991) conducted a meticulous comparison of broad and narrow transcriptions by trained personnel. For broad transcriptions, the agreement between human annotators was generally acceptable. However, for narrow transcriptions involving diacritics, the agreements were "*below acceptable reliability boundary levels, even at the least strict agreement criteria*" (Shriberg and Lof, 1991).

Given the subjectivity of phone transcriptions, we focus our efforts on **broad transcription**. Broad transcription encodes only the most salient phonetic features, usually the base vowels and consonants with infrequent use of diacritics. This is in contrast to **narrow transcription**, where the transcriber will try to transcribe as many subphonemic or phonetic details as possible with the frequent use of diacritics (Ladefoged and Johnson, 2014). Since objectively true transcriptions might not exist, we evaluate our transcriptions with phonetic feature error rates (PFER) (Taguchi et al., 2023), measuring the distance between binary articulatory features.

3 Data

First, we have created IPAPACK++, one of the largest phone-based speech corpora in 88 languages, totaling 17,132 hours. While the original IPA PACK (Zhu et al., 2024) provides 2000+ hours of speech in 100+ languages, upon careful inspection, we noticed several shortcomings. First, the IPA transcriptions were not normalized across the corpus, such that different Unicode encodings were present for the same phone. Some non-IPA Unicode symbols were also present due to artifacts in preprocessing. Second, the original dataset was more suitable for keyword spotting than ASR as

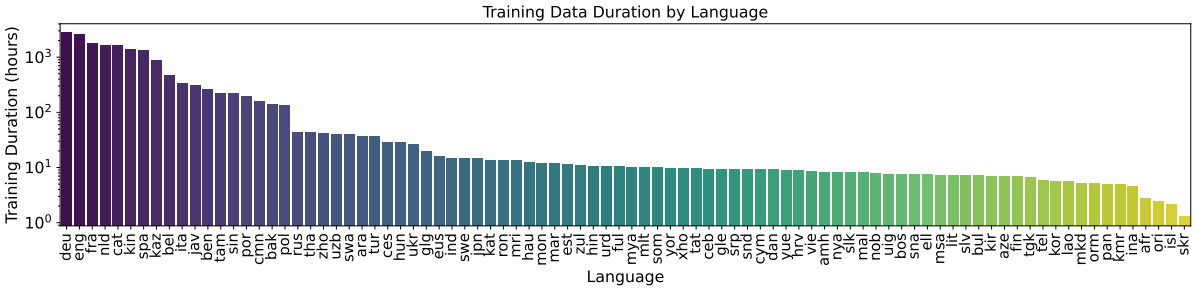

Figure 1: The distribution of labeled training data duration by language, totaling 17,132 hours.

half of the corpus was short clips of words taken from continuous recordings.

3.1 Data selection

To address some of these limitations and expand our efforts, we have created a large-scale speech dataset for phone recognition. The datasets are recreated from IPA PACK (Zhu et al., 2024), Common Voice 16.0 (Ardila et al., 2020), LibriSpeech (Panayotov et al., 2015), Multilingual LibriSpeech (Pratap et al., 2020), Aishell-1 (Bu et al., 2017), crowd-sourced speech corpora for Javanese, Sinhala, and Bengali (Kjartansson et al., 2018), IISc-MILE Tamil ASR Corpus (Madhavaraj et al., 2022b,a), Kazakh Speech Dataset (Mansurova and Kadyrbek, 2023) and Kazakh Speech Corpus (Khassanov et al., 2021). CharsiuG2P (Zhu et al., 2022) and Epitran (Mortensen et al., 2018) were used to automatically create phonemic transcriptions of available languages.

After preprocessing, we ended up with around 17,135 hours of training data with G2P-generated transcriptions in 88 languages. The language distribution of the IPAPACK++ is shown at Figure 1. A complete breakdown of individual languages can be found in Appendix A.

3.2 Tokenization

The prior state-of-the-art universal phone recognizer (Xu et al., 2022) adopted a data-driven approach to tokenization. However, this approach is not without problems. The phone tokenizer includes plain phones as well as phone combinations that are highly language-specific. For example, it uses a numerical representation of Mandarin tones rather than the standard IPA tone notations, also known as Chao tone letters. The numerical representation represents symbolic phonological contrasts of the tones, whereas the Chao tone letters reflect aspects of the phonetic realization like the f0 contour. Overall, though, using inconsistent symbols can limit knowledge sharing across languages (Zhu et al., 2024).

We further made a systematic effort to normalize IPA encodings. In the first round of filtering, PHOIBLE (Moran et al., 2014) was used as a reference to determine whether a phone was legitimate. Illegitimate phones were corrected: 1) phones with more than 3 diacritics can be overly complex to transcribe, so they are simplified to no more than one; 2) phones with inconsistent Unicode encodings, such as [g] (Unicode: U+0067) and [g] (Unicode: U+0261), are unified in one representation. Since we only focused on broad transcriptions, our final tokenizer only consists of all individual IPA symbols and the 15 most frequent diacritics from the IPA chart. Each diacritic is encoded as a separate token to reduce the vocabulary size.

3.3 Evaluation set

Evaluating on seen languages We used the test set of several publicly available datasets to evaluate model performance. The G2P-generated phone transcriptions are quite noisy (Samir et al., 2024), especially for low-resource languages. Therefore, we selected the test sets from Aishell-1 (Bu et al., 2017), Librispeech (Panayotov et al., 2015) and Multilingual LibriSpeech (MLS) (Pratap et al., 2020), where the phone transcription quality was determined to be good upon our inspection.

Evaluating on unseen languages In order to test how universal phone recognition models generalize across languages, we reserved the VoxAngeles (Chodroff et al., 2024), a clean version of the UCLA Phonetic Corpus (Li et al., 2021), and DoReCo (Paschen et al., 2020) for evaluation on unseen languages. Both datasets consist of speech recordings collected from fieldwork and transcribed phonetically by trained linguists.

Dataset	Dur.	Description
Doreco	19 hrs	45 languages collected and transcribed by field linguists.
VoxAngeles	1.5 hrs	A set of individual word recordings from 95 languages.
Buckeye	8 hrs	A collection of sociolinguistic recordings, carefully annotated by trained phoneticians.
L2-Standard	4 hrs	L2-ARCTIC speech corpus with dictionary-based phonetic transcriptions.
L2-Perceived	4 hrs	L2-ARCTIC speech corpus with human transcriptions of the actual pronunciation.
Seen languages	65 hrs	Test sets from Aishell, LibriSpeech, and the Multilingual LibriSpeech (except for English).

Table 1: A list of the evaluation datasets. These datasets cover a wide range of languages and sociophonetic conditions.

Evaluating on sociophonetic variation Most phone recognition models are trained and evaluated on dictionary pronunciations generated from pronunciation dictionaries and G2P models. These training and evaluation data might not reflect the actual pronunciation in spontaneous speech. We also measure how phone recognition models can predict actual phonetic variation. Such evaluation can serve to assess whether phone recognition models are suitable for tasks like pronunciation assessment and sociophonetic transcriptions.

Here we utilize L2 ARCTIC (Zhao et al., 2018) and the Buckeye Corpus (Pitt et al., 2005), both of which contain highly variable English speech carefully transcribed by professional linguists. For the Buckeye Corpus, we segmented all recording files into individual utterances between 20 to 50 phonemes, delimited by silent intervals ($\geq$ 200ms) (Fuchs et al., 2022). For L2 ARCTIC, we used the original segmentation but generated two versions of transcriptions, one for dictionary pronunciations of the prompts and one for the perceived pronunciations annotated by linguists.

4 Method

Some prior studies in universal phone recognition leverage knowledge of the language's phonemic inventory (Li et al., 2020; Glocker et al., 2023). However, the inventory is a static, abstract description of the phonological system of a language, only capturing a limited, idealized variation of speech. Many speech variations within a language can go beyond the inventory. In many applications of phone recognition such as pathological speech assessment, pronunciation assessment, and sociophonetics, transcribing speech into phones as it is actually articulated is important. Therefore, in our proposed models, we did not directly incorporate language-specific inventory knowledge, noting that such knowledge can also be incorporated in post-processing (Xu et al., 2022).

4.1 Zipformer

Pretrained self-supervised models such as XLS-R (Babu et al., 2022) and Whisper (Radford et al., 2023) have been utilized as base models for fine-tuning in prior studies (Xu et al., 2022; Taguchi et al., 2023; Glocker et al., 2023; Samir et al., 2024). However, fine-tuning these transformer models on our large-scale dataset is prohibitively expensive with an academic computing budget. For example, Whisper pads every input utterance, regardless of their lengths, to chunks of 30 seconds, allocating many computations to padding tokens that do not contribute to inference. Moreover, its autoregressive decoding is also highly inefficient.

Instead, we adopt Zipformer (Yao et al., 2023), a transformer encoder model with U-Net style downsampling and upsampling layers (Ronneberger et al., 2015) as the base architecture. Compared to the vanilla transformers (e.g., Wav2Vec2 and XLS-R), Conformer (Gulati et al., 2020), Branchformer (Peng et al., 2022) and E-Branchformer (Kim et al., 2023), Zipformer has demonstrated superior ASR performance with less compute (Yao et al., 2023). Zipformer achieves such compute efficiency through reusing attention weights across layers, and progressively downsampling speech in the middle layers and upsampling to the output resolution in later layers.

4.2 CR-CTC

We use the Connectionist Temporal Classification (CTC) loss (Graves et al., 2006) because it enables efficiently parallelized predictions and has maintained competitive results compared to an encoder-decoder architecture (Peng et al., 2024a). Specifically, we adopted Consistency-Regularized CTC (Yao et al., 2025) for our phone recognition model.

Given a speech-transcription pair $(\mathbf{x}, \mathbf{y})$, we fit an ASR model $f(\cdot)$. For the input speech spec-

trogram $\mathbf{x}$, $\mathbf{x}^{(a)}$ and $\mathbf{x}^{(b)}$ are two different augmented views generated through SpecAugment (Park et al., 2019). Two CTC output frame-wise distributions are generated through $\mathbf{z}^{(a)} = f(\mathbf{x}^{(a)})$ and $\mathbf{z}^{(b)} = f(\mathbf{x}^{(b)})$. Then the CR-CTC loss is formulated as:

$$\mathcal{L}_{CR-CTC} = \frac{1}{2}\left(\mathcal{L}_{CTC}(\mathbf{z}^{(a)}, \mathbf{y}) + \mathcal{L}_{CTC}(\mathbf{z}^{(b)}, \mathbf{y})\right) + \alpha\mathcal{L}_{CR}(\mathbf{z}^{(a)}, \mathbf{z}^{(b)})$$

In addition to the regular CTC loss $\mathcal{L}_{CTC}$, $\mathcal{L}_{CR}$ is used to regularize the output distributions with Kullback-Leibler (KL) divergence between two frame-wise distributions at the same time step. The CR loss is defined as:

$$\mathcal{L}_{CR}(\mathbf{z}^{(a)}, \mathbf{z}^{(b)}) = \frac{1}{2}\sum_{t=1}^{T} D_{KL}\left(sg(z_t^{(b)}), z_t^{(a)}\right) + D_{KL}\left(sg(z_t^{(a)}), z_t^{(b)}\right)$$

where $sg(\cdot)$ is the stop-gradient operation and D_{KL} is the KL divergence. $\mathcal{L}_{CR}$ performs self-distillation between outputs of two different augmented input views, mitigating overfitting. It has been shown to outperform regular CTC loss and RNN-T loss (Yao et al., 2025).

We used the original CR-CTC implementation (Yao et al., 2025) with minor modifications. The output temporal resolution is 25 Hz in the original Zipformer model. Yet this resolution is too short for phone sequences, which are significantly longer than text tokens. We upsampled the output resolution to 50 Hz to present numerical errors when computing the CTC loss. We also trained two variants of CR-CTC models: ZIPA-CR-SMALL with 64M parameters and ZIPA-CR-LARGE with 300M parameters.

4.3 Transducer

We also trained Zipformer-based transducer models (Yao et al., 2023). The original RNN-T loss for tranducers is computation- and memory-intensive, so we utilized the memory-efficient pruned RNN-T loss (Kuang et al., 2022). In the transducer, we used Zipfomer as the encoder and the stateless decoder with 1D convolutional layers (Ghodsi et al., 2020). We trained two variants with non-causual attention: ZIPA-T-SMALL with 65M parameters and ZIPA-T-LARGE with 302M parameters.

4.4 Noisy student training

Prior studies have shown that noisy student training (Park et al., 2020), or training on pseudo-labels can reliably improve multilingual ASR performance (Hwang et al., 2022a,b; Ramirez et al., 2024). We generated phone pseudo-labels for two unannotated multilingual speech datasets, VoxLingua-107 and MMS ulab v2. VoxLingua-107 (Valk and Alumäe, 2021) consists of speech recordings without transcriptions from 107 languages, totaling 6,628 hours. MMS ulab v2 (Chen et al., 2024) is a 6,700-hour speech dataset in 4,023 languages, a reproduction of the original dataset for training Meta MMS (Pratap et al., 2024). As our labelled training data only included 88 unique languages, these two datasets can tremendously enrich the language diversity of our training data.

We used all four Zipformer-based phone recognition models to generate the pseudo-labels for these two multilingual corpora, and computed the pairwise Phonetic Feature Error Rate (PFER) with PanPhon (Mortensen et al., 2016). The consistencies of predictions between models were used as a heuristic to filter out bad predictions. Speech samples with an averaged pairwise PFER higher than the 80 percentile were ultimately excluded. We used pseudo-labels from ZIPA-CR-LARGE as the final transcriptions for simplicity. Finally, we obtained pseudo-labels for 11,851 hours of multilingual speech in around 4,000 languages. We continued to train the CR-CTC models by mixing both the original dataset and pseudo-labelled dataset. The loss function was formulated as below.

$$\mathcal{L}_{mixed} = \mathcal{L}_{CR-CTC} + \lambda \cdot \mathcal{L}_{CR-CTC}^{Pseudo}$$

The hyperparameter λ was set to 0.5 to downscale the weights of the noisy pseudo-labels. We adopted noisy student training to train ZIPA-CR-NS-SMALL and ZIPA-CR-NS-LARGE, both of which were initialized from pretrained checkpoints of ZIPA-CR-SMALL and ZIPA-CR-LARGE respectively.

No-Diacritic Models Our error analysis suggested that many recognition errors were associated with diacritics. During noisy student training, we also trained two variants of ZIPA-CR-SMALL and ZIPA-CR-LARGE without diacritics. We maintained the exact same training settings, but removed all diacritics from all training data. For consistency, these models were also evaluated with the same evaluation data but without diacritics.

Model	Param.	Iters	eng-c	eng-o	ger	por	fre	spa	dut	ita	cmn	Avg.
Allosaurus	11M	-	4.18	6.21	30.26	33.09	32.77	28.02	33.29	26.57	6.64	22.33
W2V2P-lv-60-ft	300M	-	4.09	4.26	18.11	23.47	28.63	7.97	27.27	8.63	6.82	14.36
W2V2P-xlsr-53-ft	300M	-	5.45	5.35	11.61	18.80	26.59	5.14	20.91	6.93	6.20	11.88
MultIPA*	300M	-	11.26	10.86	27.02	25.05	31.31	12.02	32.15	11.22	8.35	18.80
WhisperPPT	244M	-	6.36	7.39	20.40	18.29	26.85	6.89	13.29	5.52	2.03	11.89
ZIPA-T-SMALL	65M	300k	1.17	2.14	4.66	18.20	16.68	2.34	6.84	5.05	0.72	6.42
ZIPA-T-LARGE	302M	300k	0.70	1.35	3.76	6.52	5.32	1.85	5.33	8.91	0.52	3.80
ZIPA-T-SMALL	65M	500k	0.95	1.67	3.51	17.01	7.49	2.08	5.49	2.66	0.78	4.62
ZIPA-T-LARGE	302M	500k	**0.61**	**1.19**	3.38	5.96	**4.52**	**1.69**	**4.62**	**1.91**	0.44	**2.70**
ZIPA-CR-SMALL	64M	300k	2.36	3.29	14.11	20.19	18.19	4.07	9.69	8.27	1.59	9.08
ZIPA-CR-LARGE	300M	300k	1.07	1.92	3.70	21.15	5.47	2.37	5.25	2.28	0.55	4.86
ZIPA-CR-SMALL	64M	500k	1.15	2.23	3.56	18.19	6.13	2.74	7.27	8.47	0.84	5.62
ZIPA-CR-LARGE	300M	500k	0.77	1.49	3.34	7.10	4.99	2.58	5.23	2.23	0.54	3.14
ZIPA-CR-NS-SMALL	64M	700k	0.75	1.51	3.41	8.56	4.87	2.36	4.9	2.19	0.50	3.22
ZIPA-CR-NS-LARGE	300M	800k	0.66	1.29	**3.07**	**5.47**	4.53	1.98	4.86	2.23	**0.38**	2.71
*No diacritics***												
ZIPA-CR-NS-SMALL	64M	700k	0.78	1.51	3.25	8.70	4.83	2.31	4.73	2.10	0.50	3.02
ZIPA-CR-NS-LARGE	300M	780k	0.65	1.28	2.95	4.92	4.55	2.24	4.68	2.20	0.41	2.65

Table 2: Main PFER results on seen languages. *Some languages were not seen by MultiIPA. **Diacritics were removed for both training and evaluation sets, so results are not directly comparable with other models. **Notations:** T - Transducer; CR - Consistency-regularized CTC; NS - Noisy student training.

5 Experiments

5.1 Implementation

Our experiments were structured within the Next-gen Kaldi framework. We used lhotse[1] to manage data loading and augmentation, icefall[2] for training and evaluation, and k2[3] for the pruned transducer loss. The inputs to all models are the 80-dimensional Mel Frequency Cepstral Coefficients (MFCCs). We used the Scaled Adam optimizer, which was shown to work better with Zipformer than Adam (Yao et al., 2023). All models were trained from scratch with randomly initialized weights. During evaluation, the final model for each variant was the averaged model from the last 10 checkpoints. Simple greedy decoding was used to generate predictions in all conditions. We trained all small models with an A40 40G GPU and all large models with 2 A100 40G GPUs. Detailed hyperparamters are described in Appendix B.

5.2 Baselines

To contextualize the performance of the proposed model, we compared our models with several universal phone recognition models with publicly available weights.

- **Allosaurus**. Allosaurus (Li et al., 2020) is one of the earliest universal phone recognizers. The network backbone consists of bi-directional LSTM networks, and it has a shared phone output layers and language-specific allophone layers.
- **Wav2Vec2Phoneme**. Wav2Vec2Phoneme (Xu et al., 2022) is a state-of-the-art phone recognizer based on the pretrained Wav2Vec2 and XLSR-53 model (Baevski et al., 2020; Conneau et al., 2020). It was fine-tuned on 57k hours of speech data with phone transcriptions. We examined two checkpoints: W2V2P-lv-60-ft[4], which was fine-tuned from Wav2Vec2, and W2V2P-xlsr-53-ft[5], which was initialized with XLSR-53.
- **MultIPA**. MultIPA[6] (Taguchi et al., 2023) is another model based on the XLSR-53 (Conneau et al., 2020). The model was fine-tuned on relatively small but high-quality data with phone transcriptions, achieving competitive performance in phone recognition.
- **Whisper-PPT**. Whisper-PPT (Samir et al., 2024) is an autoregressive universal phone recognition model based on the pretrained Whisper-small (Radford et al., 2023). It was

[1] https://github.com/lhotse-speech/lhotse
[2] https://github.com/k2-fsa/icefall
[3] https://github.com/k2-fsa/k2
[4] https://huggingface.co/facebook/wav2vec2-lv-60-espeak-cv-ft
[5] https://huggingface.co/facebook/wav2vec2-xlsr-53-espeak-cv-ft
[6] https://huggingface.co/ctaguchi/wav2vec2-large-xlsr-japlmthufielta-ipa1000-ns

Model	Param.	Iters	Unseen		Sociophonetic			Avg.
			Doreco	VoxAngeles	L2-Standard	L2-Perceived	Buckeye	
Allosaurus	11M	-	8.89	1.35	5.21	5.72	5.04	5.24
W2V2P-lv-60-ft	300M	-	6.13	0.66	2.89	3.95	**3.85**	3.49
W2V2P-xlsr-53-ft	300M	-	5.94	**0.58**	3.69	4.09	3.97	3.65
MultIPA	300M	-	6.55	0.62	5.86	5.88	5.94	4.97
WhisperPPT	244M	-	9.31	0.91	6.44	6.65	6.80	6.02
ZIPA-T-SMALL	65M	300k	7.20	0.71	2.26	3.85	4.00	3.60
ZIPA-T-LARGE	302M	300k	7.27	0.88	1.79	3.67	3.95	3.51
ZIPA-T-SMALL	65M	500k	8.72	0.78	1.99	3.80	3.97	3.85
ZIPA-T-LARGE	302M	500k	8.05	0.88	**1.68**	**3.63**	3.94	3.63
ZIPA-CR-SMALL	64M	300k	5.97	0.78	3.42	5.13	4.58	3.97
ZIPA-CR-LARGE	300M	300k	6.90	0.83	2.15	3.71	3.91	3.50
ZIPA-CR-SMALL	64M	500k	6.02	0.65	2.54	4.60	4.71	3.70
ZIPA-CR-LARGE	300M	500k	6.37	0.77	1.87	3.69	3.93	3.32
ZIPA-CR-NS-SMALL	64M	700k	5.94	0.69	1.94	3.79	3.87	3.24
ZIPA-CR-NS-LARGE	300M	800k	**5.93**	0.75	1.75	3.67	3.92	**3.20**
*No diacritics***								
ZIPA-CR-NS-SMALL	64M	700k	5.80	0.68	1.92	3.76	3.86	3.21
ZIPA-CR-NS-LARGE	300M	780k	5.81	0.71	1.78	3.66	3.86	3.17

Table 3: Main PFER results on unseen languages and domains. **Diacritics were removed for both training and evaluation sets, so results are not directly comparable with other models. **Notations:** T - Transducer; CR - Consistency-regularized CTC; NS - Noisy student training.

fine-tuned on a selected high-quality subset of IPAPack (Zhu et al., 2024). Unlike other models, the autoregressive nature of Whisper makes it uniquely prone to repeatedly generating hallucinated substrings on occasion.

Allophant (Glocker et al., 2023) is another state-of-the-art phone recognizer based on XLS-R (Babu et al., 2022). However, Allophant relies on an existing phoneset to make predictions. Some of our evaluation datasets, such as unseen languages and L2 speech, do not have an existing phoneset in PHOIBLE, so we did not compare with Allophant.

6 Results

We evaluated model performance with the PFER, which measures the alignment of binary articulatory features. The metric was computed with PanPhon (Mortensen et al., 2016). The main results are presented in Table 2 and Table 3. Below we summarize our main findings.

ZIPA models reach state-of-the-art performance on multilingual phone recognition. We trained ZIPA variants on 17k hours of multilingual data from scratch. Even the small ZIPA models with only 64M parameters can outperform the 300M transformer baselines that have been pretrained and/or fine-tuned on much more data. For example, both FAIR-lv-60-ft and FAIR-xlsr-60-ft (Xu et al., 2022) were initialized from pretrained weights and fine-tuned on 57k labeled data. Meanwhile, the Zipformer backbone is also much more memory efficient and less computationally intensive than the vanilla transformer in XLSR series of models and Whisper (Yao et al., 2023). Our study shows that careful curation of data, including increasing data quantity and carefully normalizing the IPA labels, as well as a good choice of backbone model can yield effective improvement.

Smaller models and non-autoregressive models generalize better to unseen languages but perform worse on seen languages. Our results show that transducer models tend to outperform CTC based models on seen languages (see Table 2). Autoregressive transducers model the dependencies better than CTC models, where conditional independence between labels is learned. However, learning the causal dependencies between phones can also hurt the multilingual generalizability, as unseen languages might have a different phonological structure. Larger models also tend to overfit the training data, weakening their abilities to predict unseen languages. This is particularly evident on both Doreco and the VoxAngeles test sets.

Yet CTC models are still valuable as they are more efficient than autoregressive models and can be combined with an external alignment algorithm to generate approximate time stamps for multilingual data (Kürzinger et al., 2020; Pratap et al.,

Model	Seen Avg.	Unseen Avg.
ZIPA-CR-L		
- 800k	3.18	3.28
ZIPA-CR-NS-L		
- λ=0.5	**2.71**	3.20
- λ=1.0	2.77	**3.19**
ZIPA-CR-NS-S		
- λ=0.2	3.36	3.35
- λ=0.5	**3.22**	3.24
- λ=1.0	3.41	**3.23**

Table 4: Ablation analysis of noisy student training.

IPA	Del	IPA	Ins	IPA	Sub
ː	17225	ː	9777	a → ɑ	5669
ʔ	11105	c̪*	3924	i → e	5592
i	7172	a̪	3401	ɛ → e	4454
a	5631	j	2793	o → u	3478
n	4714	i	2491	e → i	3089
h	4204	u	2123	u → o	2889
e	3727	e	1943	ɔ → o	2678
ɛ	3646	t	1487	ɛ → a	2480
u	3399	k	1487	ə → a	2226
~	3249	ʃ	1414	e → a	1932
o	3232	ʒ	1289	b → p	1859
t	2492	n	1265	d → t	1717
ə	2291	ʊ	894	o → ɔ	1716
ɪ	2041	o	814	i → j	1619
w	1965	p	781	ɡ → k	1609
j	1938	w	759	o → a	1526
d	1811	ə	696	i → ɪ	1444
k	1647	r	667	ɛ → e	1436
ɔ	1612	d	653	r → ɾ	1429
ʰ	1505	ɾ	648	e → ɛ	1425

Table 5: Summary of **Del**etions, **Ins**ertions, and **Sub**stitution errors by ZIPA-CR-NS-L. Other ZIPA models also exhibit a similar pattern. *c denotes any consonant.

2024).

It is important to note that the evaluation metric PFER is a distance function rather than a ratio, so its magnitude tends to correlate with length. While it appears that the PFER for seen languages (Table 2) is higher than the PFER for unseen languages (Table 3), it is because the speech samples from seen languages are longer than those from unseen languages. In Table 2, some languages, especially `por` and `fre`, have consistently lower scores than other languages. This is caused by both the length of the evaluation samples and the phone set mismatch in these languages. For example, Portuguese uses [ɐ] frequently but it is often transcribed as the more crosslinguistically frequent [a]. French marks nasality with a diacritic [~], but other languages tend to use the nasal consonants. Such mismatches in the phone set pose challenges to the phone recognition system, especially for small models. Yet longer training time and more model parameters enable models to memorize these language-specific conventions better. At least for high-resource languages, ZIPA models can implicitly distinguish the language and transcribe phones accordingly.

Noisy student training brings minor but consistent improvement. ZIPA-CR-NS models can consistently improve model performance, though the improvement is minor. This is likely because the pseudo-labels on unseen languages are extremely noisy, diminishing the benefits of additional data. Our ablation analysis in Table 4 indicates that continuing to train together with pseudo-labelled data is more beneficial than continuing to train on the existing labeled data beyond 500k steps. The value of λ seems to control the trade-off between in-domain and out-of-domain performance, but the overall impact of λ is not large. Note that we only adopted a simple approach to do noisy student training, so there is still room for improvement. Further research is needed to investigate how to better exploit the massive amount of unlabelled data.

Removing diacritics can improve the match between model predictions and ground truth, especially on unseen languages, but the impact is slight. Both Table 2 and 3 suggest that the no-diacritic condition yields inconsistent and slight improvement, as the number of total symbols is reduced. Our further inspection indicates that ZIPA models tend to handle diacritics pretty well for seen languages, as the patterns in these languages are probably well memorized during training. Yet, generalizing diacritics across languages poses a much larger challenge. The largest change in score is the `Doreco` evaluation set, as it contains more diacritics than other datasets (Paschen et al., 2020).

7 Analysis

We also conducted an error analysis to understand model behaviors and present findings below.

Phone recognition models tend to smooth out the phonetic variation during inference. In Table 3, there is a systematic gap between the performance of `L2-Standard` and the `L2-Perceived` test sets. In Figure 2, given the exact same L2 speech, ZIPA predictions tend to better match the standard dictionary pronunciation than the actual pronunciation.

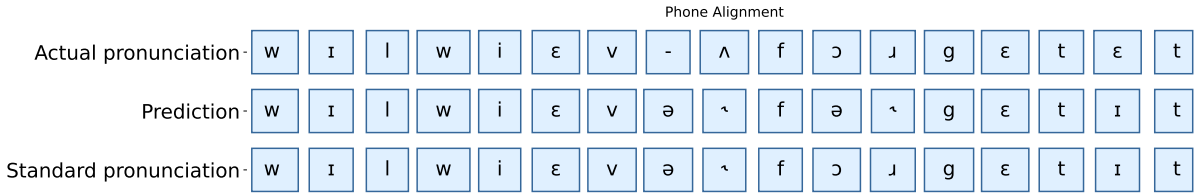

Figure 2: A sample transcription of the prompt "*Will we ever forget it*" in L2 speech by ZIPA-CR-NS-L. The predicted transcription aligns more with the standard pronunciation, suggesting that the model failed to capture the actual sociophonetic variation.

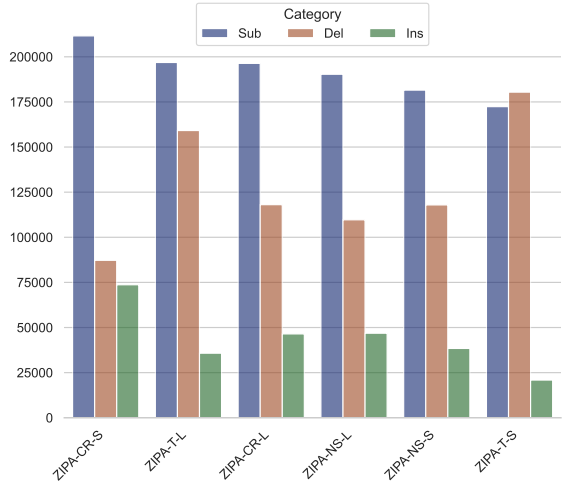

Figure 3: Distributions of transcription error types. Substitution errors are most common across models. Transducers exhibit a relatively high rate of deletion errors.

This is likely an artifact of data curation, as all of the training data were generated from pronunciation dictionaries and G2P models. Yet this finding also implies that the phone recognition models are still matching the frequent phone patterns in the dataset, rather than transcribing phones as they are actually produced.

Vowels are more difficult to predict crosslinguistically. Prior research has revealed that certain sounds are recognized better across languages (Żelasko et al., 2022). We conducted an error analysis of the model predictions on `Doreco`. As shown in Figure 3, substitution errors are far more common than addition and deletion errors. Transducer models show much higher deletion errors than CTC models. Our close inspection also suggests that transducers generate quite a few empty transcriptions for unseen languages.

Table 5 provides further details on the top errors made by ZIPA-CR-NS. The top deletion and insertion errors are diacritics. The length symbol ː is consistently the most frequently added or deleted symbol, as vowel length is relative across languages. The glottal stop ʔ is often not contrastive and not explicitly marked in IPA transcriptions, resulting in high deletions in model predictions. For substitution, the top errors are the substitution of vowels that are close in the vowel space. Compared to consonants, vowel realizations tend to be more gradient in their acoustics, resulting in higher acoustic overlap between otherwise contrastive vowel categories and therefore more ambiguous. Such misidentification patterns also mirror the patterns of human speech perception crosslinguistically (Sebastián-Gallés, 2005).

8 Conclusions

In conclusion, we present a large-scale multilingual phone recognition dataset IPA PACK++ and a series of Zipformer-based ZIPA models, which exhibit state-of-the-art performance on phone recognition. We hope that our research can provide foundations to support more downstream multilingual speech processing tasks that benefit from phonetic transcriptions. Yet simply scaling up the G2P for transcribed speech data alone might not be able to solve phone recognition, as models can simply memorize the standard pronunciation. We will also actively explore how to incorporate more linguistic knowledge to further improve performance.

Ethics statement

We adhere to ethical practices in our research. We only selected publicly available datasets with permissive licenses that allow us to redistribute the processed data and the models. We believe that open-sourcing our research can help facilitate future research towards multilingual speech technologies for both the speech processing communities and the linguistics communities.

It is our firm belief that this research can contribute to the promotion of more inclusive speech

technologies for more languages, especially for under-represented languages. While our model is primarily developed to support language documentation and other downstream applications, we are also aware that multilingual speech recognition can exhibit biases towards non-mainstream accents and potentially be used for malicious purposes such as surveillance. We urge that caution be exercised when deploying such models in downstream tasks.

Limitations

Our study is still limited in several ways. First, the number of languages studied in our paper is still limited. The distribution of languages is highly skewed in our dataset, which still biases our models towards high-resource languages.

Secondly, our current approach trains models on synthetic labels from G2P. However, the data quality is limited as dictionary pronunciations might not reflect the actual pronunciation in spontaneous speech. This also results in the ZIPA models to smooth out variation in some L2 speech. More research is needed to investigate how to curate higher quality data for phone recognition that can reflect the actual pronunciation.

The limitation of computational resources also limits our abilities to perform extensive hyperparameter tuning and conduct extensive experiments to explore different architectures and pseudo-labeling strategies. In the future, we will continue to explore better strategies to continue to improve the performance of multilingual speech processing systems.

Acknowledgments

This research was enabled in part through the computational resources provided by Advanced Research Computing at the University of British Columbia and the Digital Research Alliance of Canada. FS is supported by an NSERC PGS-D Scholarship. The research activities were also supported by the NSERC Discovery Grant and the CFI JELF Grant awarded to JZ and by SNF Grant PR00P1_208460 to EC.

References

Arthur S Abramson and Douglas H Whalen. 2017. Voice onset time (vot) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of phonetics*, 63:75–86.

Cormac Anderson, Tiago Tresoldi, Simon J Greenhill, Robert Forkel, Russell Gray, and Johann-Mattis List. 2023. Variation in phoneme inventories: quantifying the problem and improving comparability. *Journal of Language Evolution*, page lzad011.

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

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.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

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.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pages 1–5. IEEE.

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

Taehong Cho and Peter Ladefoged. 1999. Variation and universals in vot: evidence from 18 languages. *Journal of phonetics*, 27(2):207–229.

Eleanor Chodroff, Alessandra Golden, and Colin Wilson. 2019. Covariation of stop voice onset time across languages: Evidence for a universal constraint on phonetic realization. *The Journal of the Acoustical Society of America*, 145(1):EL109–EL115.

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, Torino, Italia. ELRA and ICCL.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Siyuan Feng, Ming Tu, Rui Xia, Chuanzeng Huang, and Yuxuan Wang. 2023. Language-universal phonetic encoder for low-resource speech recognition. In *Interspeech 2023*, pages 1429–1433.

Tzeviya Fuchs, Yedid Hoshen, and Yossi Keshet. 2022. Unsupervised word segmentation using k nearest neighbors. In *Interspeech 2022*, pages 4646–4650.

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*, pages 1304–1308.

Mohammadreza Ghodsi, Xiaofeng Liu, James Apfel, Rodrigo Cabrera, and Eugene Weinstein. 2020. Rnn-transducer with stateless prediction network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7049–7053. IEEE.

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

Yuan Gong, Ziyi Chen, Iek-Heng Chu, Peng Chang, and James Glass. 2022. Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7262–7266. IEEE.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040.

Dongseong Hwang, Ananya Misra, Zhouyuan Huo, Nikhil Siddhartha, Shefali Garg, David Qiu, Khe Chai Sim, Trevor Strohman, Françoise Beaufays, and Yanzhang He. 2022a. Large-scale asr domain adaptation using self-and semi-supervised learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6627–6631. IEEE.

Dongseong Hwang, Khe Chai Sim, Zhouyuan Huo, and Trevor Strohman. 2022b. Pseudo label is better than human label. In *Interspeech 2022*, pages 1421–1425.

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

Paul Kerswill and Susan Wright. 1990. The validity of phonetic transcription: Limitations of a sociolinguistic research tool. *Language variation and change*, 2(3):255–275.

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.

Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J Han, and Shinji Watanabe. 2023. E-branchformer: Branchformer with enhanced merging for speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 84–91. IEEE.

Oddur Kjartansson, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, and Linne Ha. 2018. Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 52–55, Gurugram, India.

Fangjun Kuang, Liyong Guo, Wei Kang, Long Lin, Mingshuang Luo, Zengwei Yao, and Daniel Povey. 2022. Pruned rnn-t for fast, memory-efficient asr training. In *Interspeech 2022*, pages 2068–2072.

Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. 2020. Ctc-segmentation of large corpora for german end-to-end speech recognition. In *International Conference on Speech and Computer*, pages 267–278. Springer.

Peter Ladefoged and Keith Johnson. 2014. *A course in phonetics*. Cengage learning.

William Lane and Steven Bird. 2021. Local word discovery for interactive transcription. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2058–2067.

Sang-Hoon Lee, Hyeong-Rae Noh, Woo-Jeoung Nam, and Seong-Whan Lee. 2022. Duration controllable voice conversion via phoneme-based information bottleneck. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1173–1183.

Xiaochang Li. 2017. *Divination engines: A media history of text prediction*. Ph.D. thesis, New York University.

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

Xinjian Li, David R Mortensen, Florian Metze, and Alan W Black. 2021. Multilingual phonetic dataset for low resource speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6958–6962. IEEE.

Chang Liu, Zhen-Hua Ling, and Ling-Hui Chen. 2023. Pronunciation dictionary-free multilingual speech synthesis using learned phonetic representations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yajing Liu, Xiulian Peng, Zhiwei Xiong, and Yan Lu. 2021. Phoneme-based distribution regularization for speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 726–730. IEEE.

A Madhavaraj, Bharathi Pilar, and Ramakrishnan A G. 2022a. Knowledge-driven subword grammar modeling for automatic speech recognition in tamil and kannada. *arXiv preprint*.

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

Madina Mansurova and Nurgali Kadyrbek. 2023. The development of a kazakh speech recognition model using a convolutional neural network with fixed character level filters. In *Proceedings of the Big Data and Cognitive Computing*, pages 5–9.

Steven Moran, Daniel McCloy, and Richard Wright. 2014. Phoible online.

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori 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, Osaka, Japan. The COLING 2016 Organizing Committee.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, pages 2613–2617.

Daniel S. Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V. Le. 2020. Improved noisy student training for automatic speech recognition. In *Interspeech 2020*, pages 2817–2821.

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 Twelfth Language Resources and Evaluation Conference*, pages 2657–2666.

Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe. 2022. Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding. In *International Conference on Machine Learning*, pages 17627–17643. PMLR.

Yifan Peng, Yui Sudo, Muhammad Shakeel, and Shinji Watanabe. 2024a. Owsm-ctc: An open encoder-only speech foundation model for speech recognition, translation, and language identification. *arXiv preprint arXiv:2402.12654*.

Yifan Peng, Yui Sudo, Muhammad Shakeel, and Shinji Watanabe. 2024b. OWSM-CTC: An open encoder-only speech foundation model for speech recognition, translation, and language identification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10192–10209, Bangkok, Thailand. Association for Computational Linguistics.

Jan Pirklbauer, Marvin Sach, Kristoff Fluyt, Wouter Tirry, Wafaa Wardah, Sebastian Moeller, and Tim Fingscheidt. 2023. Evaluation metrics for generative speech enhancement methods: Issues and perspectives. In *Speech Communication; 15th ITG Conference*, pages 265–269. VDE.

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, et al. 2024. Scaling speech technology to 1,000+

languages. *Journal of Machine Learning Research*, 25(97):1–52.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, pages 2757–2761.

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.

Francis McCann Ramirez, Luka Chkhetiani, Andrew Ehrenberg, Robert McHardy, Rami Botros, Yash Khare, Andrea Vanzo, Taufiquzzaman Peyash, Gabriel Oexle, Michael Liang, et al. 2024. Anatomy of industrial scale multilingual asr. *arXiv preprint arXiv:2404.09841*.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

Elizabeth Salesky, Eleanor Chodroff, Tiago Pimentel, Matthew Wiesner, Ryan Cotterell, Alan W Black, and Jason Eisner. 2020. A corpus for large-scale phonetic typology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4526–4546, Online. Association for Computational Linguistics.

Farhan Samir, Emily P Ahn, Shreya Prakash, Márton Soskuthy, Vered Shwartz, and Jian Zhu. 2024. Efficiently identifying low-quality language subsets in multilingual datasets: A case study on a large-scale multilingual audio dataset. *arXiv preprint arXiv:2410.04292*.

Núria Sebastián-Gallés. 2005. Cross-language speech perception. *The handbook of speech perception*, pages 546–566.

Siyuan Shan, Yang Li, Amartya Banerjee, and Junier B Oliva. 2024. Phoneme hallucinator: One-shot voice conversion via set expansion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14910–14918.

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

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.

Jörgen Valk and Tanel Alumäe. 2021. Voxlingua107: a dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658. IEEE.

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. 2023. Zipformer: A faster and better encoder for automatic speech recognition. In *The Twelfth 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. 2021. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. In *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. *Interspeech 2018*.

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 *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 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 *Proc. Interspeech 2022*, pages 446–450.

A Dataset details

A.1 Dataset Overview

Language	Split	Dataset	Total Duration
swa	train	Common Voice	28:48:36
		Fleurs	10:06:09
spa	train	Common Voice	404:00:29
		Fleurs	06:43:33
		Multilingual Librispeech	917:41:03
bel	train	Common Voice	452:08:22
		Fleurs	07:18:15
tam	train	Common Voice	76:47:49
		Fleurs	06:20:35
		IISc-MILE Tamil ASR Corpus	133:17:27
kin	train	Common Voice	1376:18:20
eng	train	Common Voice	1584:30:15
		Librispeech	961:03:15
		Fleurs	05:38:28
ron	train	Common Voice	05:44:05
		Fleurs	07:38:47
ell	train	Fleurs	07:30:24
jpn	train	Fleurs	05:03:22
		Common Voice	09:22:26
tur	train	Fleurs	06:25:40
		Common Voice	29:30:44
hun	train	Common Voice	21:15:06
		Fleurs	07:00:41
mon	train	Fleurs	08:37:49
		Common Voice	03:13:37
ind	train	Common Voice	07:46:52
		Fleurs	06:56:13
uig	train	Common Voice	07:36:34
ita	train	Common Voice	83:14:54
		Fleurs	06:51:31
		Multilingual Librispeech	247:22:40
mkd	train	Fleurs	05:08:08
urd	train	Common Voice	04:58:30
		Fleurs	05:20:32
vie	train	Fleurs	06:42:36
		Common Voice	01:51:31
cat	train	Common Voice	1591:25:03
		Fleurs	05:46:13
fra	train	Common Voice	661:14:43
		Multilingual Librispeech	1076:34:49
mya	train	Fleurs	10:04:25
kaz	train	Kazakh Speech Dataset	554:47:31
		Kazakh Speech Corpus	318:25:26
		Fleurs	08:54:35
deu	train	Common Voice	778:17:18
		Multilingual Librispeech	1966:30:30
		Fleurs	06:52:51
kir	train	Fleurs	06:59:30
mlt	train	Fleurs	07:29:59

		Common Voice	02:24:53
bos	train	Fleurs	07:34:14
srp	train	Common Voice	01:08:08
		Fleurs	08:08:18
isl	train	Fleurs	02:06:30
ori	train	Fleurs	02:25:20
pol	train	Fleurs	07:13:49
		Multilingual Librispeech	103:38:57
		Common Voice	24:47:44
nld	train	Common Voice	38:07:31
		Fleurs	05:48:46
		Multilingual Librispeech	1554:14:38
slv	train	Fleurs	05:46:47
		Common Voice	01:24:43
tel	train	Fleurs	05:52:07
hin	train	Common Voice	05:17:16
		Fleurs	05:08:23
ukr	train	Fleurs	06:41:46
		Common Voice	19:56:08
yor	train	Common Voice	01:20:01
		Fleurs	08:27:42
aze	train	Fleurs	06:53:37
zho	train	Common Voice	42:04:06
mri	train	Fleurs	13:20:08
rus	train	Fleurs	06:16:41
		Common Voice	37:26:56
swe	train	Common Voice	08:10:51
		Fleurs	06:20:35
pan	train	Fleurs	04:57:37
mar	train	Common Voice	02:13:16
		Fleurs	09:28:59
dan	train	Fleurs	05:45:06
		Common Voice	03:16:57
zul	train	Fleurs	11:03:07
nob	train	Fleurs	07:57:37
por	train	Common Voice	22:38:41
		Multilingual Librispeech	160:57:47
		Fleurs	07:45:54
ben	train	Crowd-sourced speech for Bengali	215:24:21
		Common Voice	31:49:44
		Fleurs	08:10:49
bak	train	Common Voice	139:12:22
amh	train	Fleurs	08:15:36
est	train	Fleurs	05:22:55
		Common Voice	05:49:26
cmn	train	Aishell-1	150:50:14
		Fleurs	06:02:12
ces	train	Fleurs	06:22:34
		Common Voice	22:25:29
snd	train	Fleurs	09:08:45
glg	train	Fleurs	05:07:12
		Common Voice	14:01:47

uzb	train	Common Voice Fleurs	32:39:44 07:35:51
nya	train	Fleurs	08:13:52
tat	train	Common Voice	09:29:35
kor	train	Fleurs	05:40:36
gle	train	Fleurs	09:18:51
eus	train	Common Voice	15:56:07
orm	train	Fleurs	05:06:30
mal	train	Common Voice Fleurs	00:36:24 07:22:11
ara	train	Fleurs Common Voice	04:56:05 31:58:14
slk	train	Common Voice Fleurs	03:26:03 04:32:55
hau	train	Common Voice Fleurs	02:06:03 10:05:18
yue	train	Common Voice Fleurs	03:26:30 05:33:36
ceb	train	Fleurs	09:19:35
tha	train	Fleurs Common Voice	06:12:42 37:07:21
ful	train	Fleurs	10:16:26
afr	train	Fleurs	02:42:43
kat	train	Common Voice Fleurs	09:34:08 03:52:10
fin	train	Fleurs	06:44:46
tgk	train	Fleurs	06:31:01
lit	train	Fleurs	07:16:38
sin	train	Crowd-sourced speech for Sinhala	215:47:11
cym	train	Fleurs	09:07:12
kmr	train	Common Voice	04:55:01
msa	train	Fleurs	07:17:01
jav	train	Crowd-sourced speech for Javanese Fleurs	295:46:56 08:36:13
xho	train	Fleurs	09:46:42
bul	train	Fleurs	07:02:45
ina	train	Common Voice	04:32:09
skr	train	Common Voice	01:17:07
hrv	train	Fleurs	08:46:37
sna	train	Fleurs	07:33:33
som	train	Fleurs	09:50:14
lao	train	Fleurs	05:34:58

Table 6: Detailed statistics of IPAPack++. Only the train split of the original datasets were kept. Each language is represented by the ISO 639-3 standard code.

The detailed breakdown of VoxAngeles can is available at Chodroff et al. (2024) and the detailed descriptions of Dorecos-IPA can be found at Zhu et al. (2024). The full breakdown of individual languages is listed at Table 6.

A.2 Final training data

For final training data, we removed low quality samples based on the following criteria.

- Audio samples longer than 24 seconds or shorter 1 second, which account for less than 0.01% of samples.
- IPA sequences longer than 512 tokens or shorter than 5 tokens, as determined by the tokenizer.
- IPA sequences longer than 90% of the output frame length, which can lead to inf loss values for CTC models. The 90% ratio also accounts for the speed perturbation.

All data were partitioned into individual shards of 20,000 samples using the `shar` format in `lhotse`. All shards were randomly shuffled during model training. The detailed statistics can be found in Table 7.

A.3 Pseudo-labeled data

For the VoxLingua-107 (Valk and Alumäe, 2021), we used the original segmented sentences. For the MMS ulab V2 (Peng et al., 2024b), the original audios were not segmented. We also failed to apply voice activity detection due to the presence of background noises and music. So we randomly segmented the audio into individual chunks by uniformly sampling the chunk length between 1 and 20 seconds.

Same as the original training data, all pseudo-labelled data were also partitioned into individual shards of 20,000 samples using the `shar` format in `lhotse`. The detailed statistics can be found in Table 8.

B Training details

All hyperparameters for model training are presented in Table 9 and 10. Unless otherwise stated, we adopted the original hyperparameters in the Zipformer recipe in Icefall[7]. For noisy student training, we initialized the model with the latest ZIPA-CR checkpoints at 500k steps for both sizes and continued to train the model by mixing the labeled data and the pseudo-labeled data at each step.

[7] https://github.com/k2-fsa/icefall/tree/master/egs/librispeech/ASR/zipformer

Audio count:	8,289,886
Total duration (hh:mm:ss)	17132:58:48
mean	7.4
std	4.4
min	1.0
25%	4.2
50%	5.7
75%	8.7
99%	19.7
99.5%	20.0
99.9%	20.0
max	24.0

Table 7: Summary Statistics of the final labeled training data

Audio count:	4,270,280
Total duration (hh:mm:ss)	11,851:31:53
mean	10.0
std	4.6
min	1.0
25%	6.0
50%	9.0
75%	13.2
99%	20.0
99.5%	20.0
99.9%	20.0
max	20.0

Table 8: Summary Statistics of the pseudo-labeled training data

Hyperparameter	ZIPA-T-SMALL	ZIPA-T-LARGE
Feedforward Dimensions	512,768,1024,1536,1024,768	768,768,1536,2048,1536,768
Encoder Dimensions	192,256,384,512,384,256	512,512,768,1024,768,512
Num. Layers	2,2,3,4,3,2	4,3,4,5,4,4
Downsampling factors	1,2,4,8,4,2	
Output downsampling factor	4	
Joiner dimension	512	1024
Decoder dimension	512	1024
Parameters	65M	302M
Initial Learning Rate	0.035	0.025
Optimizer	Scaled Adam	
Scheduler	Eden Scheduler	
Total Training Steps	500k	
Effective Batch Size	800 seconds	600 seconds
Mixed Precision	bfloat16	
GPUs	A40 48G	2 $\times$ A100 40G
Training time	5 days	4 days

Table 9: Hyperparameters for ZIPA-T models.

Hyperparameter	ZIPA-CR-SMALL	ZIPA-CR-LARGE
Feedforward Dimensions	512,768,1024,1536,1024,768	768,768,1536,2048,1536,768
Encoder Dimensions	192,256,384,512,384,256	512,512,768,1024,768,512
Num. Layers	2,2,3,4,3,2	4,3,4,5,4,4
Downsampling factors	1,2,4,8,4,2	
Output downsampling factor	2	
Parameters	64M	300M
Initial Learning Rate	0.035	0.025
SpecAug: Num. frame masks	20	
SpecAug: Max mask fraction	0.3	
Optimizer	Scaled Adam	
Scheduler	Eden Scheduler	
Total Training Steps	500k	
Effective Batch Size	500 seconds	240 seconds
Mixed Precision	bfloat16	
GPUs	A40 48G	2 $\times$ A100 40G
Training time	6 days	4 days
Noisy student training	ZIPA-CR-NS-SMALL	ZIPA-CR-NS-LARGE
Steps	200k	280k
Training time	5 days	4 days
Initial learning rate	1e-3	

Table 10: Hyperparameters for ZIPA-T models.