The taste of IPA 🍻: Towards open-vocabulary keyword spotting and forced alignment in any language

Jian Zhu[β,œ] **Changbing Yang**[β,œ] **Farhan Samir**[β,œ] **Jahurul Islam**[β]
[β]Department of Linguistics, University of British Columbia
[œ] Natural Language Processing Group, University of British Columbia
jian.zhu@ubc.ca {cyang33,fsamir}@mail.ubc.ca

Abstract

In this project, we demonstrate that phoneme-based models for speech processing can achieve strong crosslinguistic generalizability to unseen languages. We curated the IPAPACK, a massively multilingual speech corpora with phonemic transcriptions, encompassing 115 languages from diverse language families, selectively checked by linguists. Based on the IPAPACK, we propose CLAP-IPA, a multilingual phoneme-speech contrastive embedding model capable of open-vocabulary matching between arbitrary speech signals and phonemic sequences. The proposed model was tested on 95 unseen languages, showing strong generalizability across languages. Temporal alignments between phonemes and speech signals also emerged from contrastive training, enabling zeroshot forced alignment in unseen languages. We further introduced a neural forced aligner IPA-ALIGNER by finetuning CLAP-IPA with the Forward-Sum loss to learn better phone-to-audio alignment. Evaluation results suggest that IPA-ALIGNER can generalize to unseen languages without adaptation.

1 Introduction

The diversity of human speech presents a formidable challenge to multilingual speech processing systems. Recently, accumulating evidence indicates that scaling up the multilingual data can tremendously improve the performance of multilingual speech processing (Conneau et al., 2020; Babu et al., 2021; Radford et al., 2023; Pratap et al., 2023). However, it remains incredibly difficult, if not impossible, to gather large-scale data from every language in the world. It is becoming increasingly critical to develop speech processing systems that generalize to arbitrary unseen languages.

Despite the seeming diversity, sounds of human speech are highly constrained by the anatomical structure of the human vocal tract, which is universally shared by all humans (Gick et al., 2013). Typological research has also shown that most, if not all, human speech can be represented by around 150 phonemes and diacritics (Moran et al., 2014; Gordon, 2016). The limited degrees of freedom in human articulation have enabled phoneticians and linguists to craft universal symbolic representations of human speech, that is, the **International Phonetic Alphabet (IPA)** (International Phonetic Association, 1999).

Prior studies have shown that phoneme-based ASR models generalize to unseen languages (Li et al., 2020; Xu et al., 2022; Glocker et al., 2023). In this project, we aim to provide yet another positive answer to this central question: **can we build multilingual speech processing systems that generalize to arbitrary languages through the use of universal IPA symbols?** Specifically, we focus on two classic tasks in speech processing, **key word spotting (KWS)** and **forced alignment**. KWS is a task of identifying specific keywords in streaming speech, whereas forced alignment refers to aligning intervals of a speech signal to a given sequence of phonetic symbols. Both tasks are relevant in many practical applications such as voice assistant, speech synthesis, language documentation, etc. Yet neither task has been tackled with general systems that generalize to all languages.

This study represents an attempt to build crosslinguistically generalizable systems for KWS and forced alignment. First, we present the IPAPACK, a multilingual speech corpora in 115 languages with phonemic transcriptions, totaling over 1000 hours and carefully checked by trained linguists. Secondly, with the IPAPACK, we proposed Contrastive Language-Audio Pretraining with International Phonetic Alphabet (CLAP-IPA), a phoneme-to-speech retrieval model with contrastive pretraining on phoneme-speech pairs. Evaluations on 95 unseen languages suggest that CLAP-IPA is capable of performing zero-shot open-vocabulary KWS in any language without adaption, including lan-

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
June 16-21, 2024 ©2024 Association for Computational Linguistics

guages not seen during training.

Thirdly, we also introduce a multilingual forced alignment model, IPA-ALIGNER, that works for arbitrary languages. We noticed that alignments between phonemes and speech signals emerge from CLAP-IPA, even with only sequence-level contrastive learning. Crosslinguistic zero-shot forced alignment can be achieved with CLAP-IPA. After finetuning CLAP-IPA with an alignment loss, we propose IPA-ALIGNER that can provide crosslinguistic word-level and phone-level alignment generalizable to unseen languages. Finally, our analysis indicates that phonemes, being shared across all languages, enhance knowledge transfer within training data, serving as more effective modeling units than texts in current multilingual tasks.

We envision that our dataset and models will benefit more downstream tasks and applications in multilingual speech processing. To facilitate future research, we will release our dataset, scripts, and pre-trained models at: `https://github.com/lingjzhu/clap-ipa`.

2 Backgrounds

2.1 Spoken keyword detection and retrieval

Most research in keyword spotting focuses predominantly on English (e.g., Chen et al., 2014; Tang and Lin, 2018; Rybakov et al., 2020; Berg et al., 2021). In recent years, there has been increased interest in building multilingual keyword detection systems that can adapt to new words or new languages through few-shot learning (Mazumder et al., 2021a; Lei et al., 2023; Reuter et al., 2023). While texts are the primary modeling units in most systems, studies are showing the effectiveness of using IPA symbols to achieve open-vocabulary generalization (Tanaka et al., 2001; Shin et al., 2022; Lee and Cho, 2023; Reuter et al., 2023).

Another approach for keyword matching is based on contrastive learning frameworks, notably CLAP (Wu et al., 2023) and the subsequent CLARA (Noriy et al., 2023). Contrastive learning also enables keyword retrieval systems based on semantics rather than the surface acoustic form (Duquenne et al., 2021; Khurana et al., 2022; Zhu et al., 2022a). The contrastive learning paradigm has also been applied successfully to build open-vocabulary KWS systems (Nishu et al., 2023).

Nevertheless, existing multilingual KWS systems face limitations in terms of limited supported languages, and cannot achieve zero-shot adaptation. Built on these prior efforts, we scaled up the phoneme-based open-vocabulary KWS models to more languages to achieve crosslinguistic generalization.

2.2 Forced alignment

Forced alignment is another classic task in speech processing for segmenting speech into utterances, words, or phonemes. It is widely used for downstream tasks where phone or word durations are needed, including speech synthesis, speech assessment, language documentation, and speech corpora construction. Currently, some of the most popular forced alignment systems are still based on Hidden Markov Models (HMM), including the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017), WebMAUS (Kisler et al., 2012) and Forced Alignment and Vowel Extraction (FAVE) (Rosenfelder et al., 2011). Recently, since neural networks gradually dominate speech processing, research in performing forced alignment with deep neural models is also gaining momentum (Kelley and Tucker, 2018; Kürzinger et al., 2020; Schulze-Forster et al., 2020; Teytaut and Roebel, 2021; Teytaut et al., 2022; Zhu et al., 2022c). Neural models usually exhibit stronger performance over HMM-based systems. However, forced alignment systems are mostly set up to work in monolingual settings. Scant attention has been paid to the building of multilingual forced alignment systems that can work for multilingual languages simultaneously.

3 Dataset curation

Most speech corpora are distributed as audio-text pairs. In comparison, phonemically transcribed speech corpora are rare. Unlike text transcription, transcribing speech signals into IPA, or phonemic transcriptions often require years of expertise in phonetics, making it hard to create high-quality phonemic datasets at scale. However, these IPA symbols provide a universal representation of speech sounds such that any language can be transcribed symbolically. So IPA symbols can be used as a proxy to train multilingual speech processing systems. As a first step, we created large-scale phonemic transcriptions for public speech corpora, encompassing 115 languages across language families. The transcription can be automated through **grapheme-to-phoneme conversion (G2P)**, a process of converting orthographic transcriptions into phonemic transcriptions through pronunciation dictionaries and/or statistical models.

	Train (hrs)	Dev (hrs)	Test (hrs)	Total (hrs)	Languages	Avg. Dur (hrs)
VoxCommunis (Ahn and Chodroff, 2022)	803.84	-	-	803.84	38	21.15
IPAPACK						
FLEURS-IPA	544.02	73.46	162.06	779.54	77	10.12
MSWC-IPA	485.35	64.08	64.11	613.44	36	17.04
DORECO-IPA	13.70	-	5.29	18.99	44	0.44

Table 1: Descriptive statistics of the IPAPACK and a selected subset of VoxCommunis (Ahn and Chodroff, 2022).

3.1 Phonemic transcriptions

We primarily made use of three existing multilingual speech datasets, FLEURS (Conneau et al., 2023), Multilingual Spoken Words Corpus (MSWC) (Mazumder et al., 2021b) and DoReCo (Paschen et al., 2020).

FLEURS We used two multilingual G2P systems, Epitran (Mortensen et al., 2018) and CharsiuG2P (Zhu et al., 2022b), to create phonemic transcriptions. As these two systems cover an overlapping but slightly different set of languages, combining them allowed us to maximize the diversity of languages. Before preprocessing, we removed any texts with Arabic numbers or code-switching, as G2P systems cannot process them correctly.

Yet some Asian languages do not explicitly mark word boundaries with spaces. For Mandarin Chinese, G2PW (Chen et al., 2022) was used to create the Pinyin romanizations, which were then mapped to IPA symbols. For Thai, we used PyThaiNLP (Phatthiyaphaibun et al., 2016) to perform word segmentation and G2P. For Japanese, the word segmentation was first performed with Fugashi (McCann, 2020) before G2P was applied.

MSWC As MSWC is a word-level speech corpus, creating phonemic transcriptions was straightforward. CharsiuG2P and Epitran were deployed to transcribe the orthographic words to phonemic sequences. To strike a balance between diversity and quantity, we limited the maximum frequency to 50 to prevent high-frequency words from dominating the dataset. For words with more than 50 samples, only 50 of them will be randomly selected from the pool. After filtering, we ended up with 2.3 million spoken words, amounting to around 613 hours.

DoReCo The original DoReCo data were distributed as hour-long recordings, so we segmented them into individual utterances based on the sentence boundaries in the provided annotations. For DoReCo, all languages were transcribed as phonemes using X-SAMPA (Wells, 1995) notations. We simply converted the X-SAMPA transcription to IPA symbols, as there is a one-to-one mapping between these two systems. Utterances with incomplete transcriptions or loud background noises were discarded.

3.2 Dataset validation

As G2P systems are based on rules or pronunciation dictionaries, they reflect how a word **should** be pronounced rather than how a word **is** pronounced. Given the high variability (e.g., phonetic reduction, coarticulation) in speech signals, it is not always possible for the G2P phonemic transcriptions to exactly match the audio. We were aware that a true transcription does not always exist for every utterance (Ladefoged and Halle, 1988; Ladefoged, 1990). Even trained phoneticians often disagree on the phonemic transcriptions of the same utterance, due to factors including psycho-acoustic constraints, phonetic training, and their mother tongue (Pitt et al., 2005; Heselwood, 2013).

Two authors (trained phoneticians) listened to at least ten random samples in each language to determine the transcription quality. We applied a relatively relaxed standard for the generated transcriptions: as long as the speech signal approximately matches more than 80% of the transcription, it is considered valid. While we made our best efforts to validate the transcription quality, we acknowledge that there are still transcription errors in the dataset. A summary of the IPAPACK is presented in Table 1. To augment our current dataset, we also included a filtered subset of VoxCommuis Corpus (Ahn and Chodroff, 2022), which is a multilingual speech corpora created in a similar workflow, though with slightly different pronunciation dictionaries and G2P tools. Detailed information on individual languages of the VoxCommuis Corpus is at Appendix A

4 Method

4.1 Contrastive learning for KWS

Here we adopt the same contrastive learning framework as CLAP (Wu et al., 2023), as it has been proven to be one of the most effective strategies for learning high-quality cross-modal representations. There are two separate encoders to process phoneme sequence $\mathbf{P} \in \mathbb{R}^{N\times 1}$ and speech MFCC features $\mathbf{S} \in \mathbb{R}^{T\times K}$, transforming them into phoneme embedding and speech embedding. In this study, we use the **SigLIP loss**, a simpler sigmoid-based loss that is shown to be as effective as the softmax-based CLIP loss (Zhai et al., 2023). Given two normalized embeddings $\boldsymbol{x}_i \in \mathbb{R}^D = f_S(\mathbf{P}_i)$ and $\boldsymbol{y}_i \in \mathbb{R}^D = f_T(\mathbf{S}_i)$, it is defined as follows.

$$\mathcal{L} = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}}}_{\mathcal{L}_{ij}} \quad (1)$$

where t and b are learnable parameters that were updated during training. z_{ij} is the ground truth label, $z_{ij} = 1$ for positive pairs and $z_{ij} = -1$ for negative pairs. Following the recommendation by Zhai et al. (2023), we initialized $t = \log 10$ and $b = -10$.

Speech encoder The speech encoder has the same transformer encoder architecture as the Whisper's encoder. The weights were initialized with Whisper's pre-trained encoder weights, whereas the decoder was discarded. The original Whisper encoder does not accept attention masks, but these padding tokens can bias the model during pooling. So attention masks were also passed to the speech encoder and the final fixed-dimensional embedding through mean pooling on non-padded hidden states. For speech data augmentation, SpecAugment (Park et al., 2019) was applied during training using the default hyperparameters in Whisper training.

Phoneme tokenizer We trained a specialized tokenizer to encode all base IPA symbols and diacritics, including tonal notations, stress marks, and tie bars for affricates. Upon inspection, we noticed that the IPA transcriptions were inconsistent across languages. For example, tie bars were inconsistently labeled (e.g., [tʃ] vs. [t͡ʃ]) and stress marks tend to be a language-specific phenomenon (Gordon and Roettger, 2017). Yet we did not perform normalization on these idiosyncratic labels to preserve the diversity of our data. The phoneme tokenizer was trained using the unigram algorithm (Kudo, 2018) with `sentencepiece` package [1]. The tokenizer was trained on all phonemic transcriptions in our datasets, with a vocabulary of 450 and byte-fallback for unknown characters.

Phoneme encoder For the phoneme encoder, we used the BERT architecture (Devlin et al., 2019) with mean pooling of the final hidden states as the fixed-dimensional representation. The phoneme encoder was pre-trained on a corpus of phonemic transcriptions using standard masked language modeling (MLM) as detailed in (Devlin et al., 2019). Given that phoneme sequences are of less complexity than texts, the masking probability was set to 30%. The training data were pooled from diverse sources, including the IPAPACK, pronunciation dictionaries in CharsiuG2P (Zhu et al., 2022b), and Vox Communis (Ahn and Chodroff, 2022). The final pretraining corpus consists of 11 million samples in more than 110 languages. We pre-trained three phoneme encoders of different sizes, matching hyperparameters including the number of layers, hidden dimensions, and the number of attention heads to the corresponding Whisper encoder (`tiny`, `base` and `small`).

4.2 Forced alignment

We noticed that phoneme-to-speech alignment emerged from CLAP-IPA on the pairwise cosine similarity matrix computed with the token-wise hidden states of phone and speech encoders. We introduce a simple algorithm to derive the alignment between phonetic units and speech signals, with control over the temporal resolution of speech frames and the granularity of phonetic sequences.

Adaptive average pooling While we expect the forced aligned units to be natural phonetic units like phonemes and words, due to tokenization, the hidden states of phone encoders correspond to a character or byte unit rather than a natural phonetic unit. A sequence of phonemes or words of length N' might be tokenized into a character or byte sequence of length $N, N \geq N'$. We define an adaptive average-pooling mask $\mathbf{M}_p \in \mathbb{R}^{N'\times N}$ to downsample the hidden representations. Through this pooling mask, consecutive hidden states belonging to one phoneme or one word were averaged to one fixed dimensional vector, such that each output hidden state after pooling corresponds to a natural phonetic unit (see Fig 1). This ensures that our forced alignment algorithm works for any level of phonetic units.

We can also define a similar adaptive average-

[1] https://github.com/google/sentencepiece

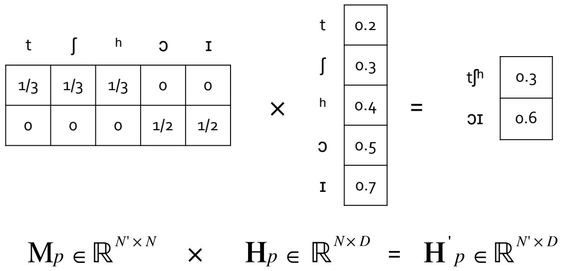
t ʃ ʰ ɔ ɪ
1/3 1/3 1/3 0 0
0 0 0 1/2 1/2
×
t 0.2
ʃ 0.3
ʰ 0.4
ɔ 0.5
ɪ 0.7
=
tʃʰ 0.3
ɔɪ 0.6
$\mathbf{M}_p \in \mathbb{R}^{N' \times N}$ × $\mathbf{H}_p \in \mathbb{R}^{N \times D}$ = $\mathbf{H}'_p \in \mathbb{R}^{N' \times D}$

Figure 1: Illustration of adaptive average-pooling of phoneme representations, $\mathbf{M_p H_p} = \mathbf{H'_p}$.

pooling mask for speech representations $\mathbf{M}_s \in \mathbb{R}^{T' \times T}$ to downsample them from length T to T'. For word-level alignments that don't require high temporal resolution, we can compress the length of the speech hidden states by controlling the pooling window length and frameshift.

Zeroshot forced alignment Given two sequences of hidden states $\mathbf{H}_s \in \mathbb{R}^{T \times D}$ and $\mathbf{H}_p \in \mathbb{R}^{N \times D}$ produced by the speech encoder and phone encoders, adaptive average-pooling masks $\mathbf{M}_p \in \mathbb{R}^{N' \times N}$ and $\mathbf{M}_s \in \mathbb{R}^{T' \times T}$ are used to transform them into more compact representations $\mathbf{H'_s} \in \mathbb{R}^{T' \times D}$ and $\mathbf{H'_p} \in \mathbb{R}^{N' \times D}$.

$$\mathbf{H'_s} = \text{Normalize}(\mathbf{M_s H_s}\text{,dim=-1})$$

$$\mathbf{H'_p} = \text{Normalize}(\mathbf{M_p H_p}\text{,dim=-1})$$

$$\mathbf{D} = \mathbf{H'_s} \mathbf{H'^{\top}_p} / \tau$$

where τ is the fixed temperature parameter and was set to 0.05 by default. The pairwise similarity matrix $\mathbf{D} \in \mathbb{R}^{T' \times N'}$ is used to derive the temporal monotonic alignment between phonetic units and speech frames through **dynamic time warping (DTW)**, even if CLAP-IPA had never between trained on alignment labels.

Finetuning To further enhance the performance of forced alignment, we introduce IPA-ALIGNER by finetuning CLAP-IPA with the **Forward-Sum Loss**, which has been shown to be effective in learning monotonic alignments between speech and phonemes (Shih et al., 2021; Badlani et al., 2022; Zhu et al., 2022c).

$$\mathcal{L} = \mathcal{L}_{ForwardSum}(\mathbf{D})$$

This alignment learning loss function relies on the forward-sum algorithm in classic HMMs to maximize the likelihood of text sequence given speech sequences, while enforcing the monotonic constraint of alignment (see Shih et al. (2021) for detailed derivations). The Forward-Sum loss requires a good prior alignment to converge to meaningful results, so we did not report failure results from randomly initialized models.

During finetuning, we only average-pooled the phoneme representations at the phoneme and kept the original speech representations (by setting $\mathbf{M}_s$ to the identity matrix $\mathbf{I}$). In inference, for phoneme alignment, we pooled the phoneme representations at the phoneme-level and kept the original speech representations. For word alignment, the phoneme representations were pooled at the word-level and the speech representations were average-pooled with a window length of 3 and a frameshift of 2.

5 Experiments

5.1 Training details

We trained three variants of models, CLAP-IPA-tiny, CLAP-IPA-base and CLAP-IPA-small, all of them were matched to the default encoder parameters of Whisper (Radford et al., 2023). The speech encoder and phoneme encoder were symmetric. Our training dataset included the training set of IPAPACK plus the VoxCommunis speech corpora (Ahn and Chodroff, 2022). By default, all models were trained with paired speech recordings and their phonemic transcriptions. For IPA-ALIGNER, we finetuned CLAP-IPA-tiny, CLAP-IPA-base and CLAP-IPA-small on the same data excluding MSWC-IPA. All detailed hyperparameters can be found in Appendix B.

For controlled comparison, we also trained two base models, CLAP-IPA-TEXT and CLAP-IPA-PHONE on the same FLEURS-IPA and MSWC-IPA subset either with only phonemic or text transcriptions. These two models were matched in total parameters, training data, and all other hyperparameters during training. In another controlled experiment, we trained CLAP-IPA-FLEURS and CLAP-IPA-VC either only on the FLEURS-IPA or the VoxCommunis, which would allow us to examine the impact of data size and language diversity.

5.2 Evaluation datasets

We evaluated the crosslinguistic generalizability of our models on several evaluation datasets covering a wide range of topologically diverse languages. Whenever possible, we made our best effort to include baseline models to contextualize our model performance. This was not always possible, because evaluating multilingual KWS and multilingual forced alignment on unseen languages are new tasks and in some cases we were not able to find

Method	LibriPhrase-Easy		LibriPhrase-Hard	
	EER(%) ↓	AUC(%) ↑	EER(%) ↓	AUC(%) ↑
CMCD (Shin et al., 2022)	8.42	96.7	32.90	73.58
PhonMatchNet (Lee and Cho, 2023)	2.80	99.29	18.82	88.52
CED (Nishu et al., 2023)	1.7	99.84	**14.4**	**92.7**
CLAP-IPA-TEXT	6.0	98.31	31.14	74.8
CLAP-IPA-PHONE	1.3	99.88	23.03	84.58
CLAP-IPA-FLEURS	0.95	99.94	22.98	84.82
CLAP-IPA-VC	0.81	99.55	21.55	85.91
CLAP-IPA-tiny	0.68	99.96	20.85	86.58
CLAP-IPA-base	0.63	99.97	20.04	88.25
CLAP-IPA-small	**0.56**	**99.97**	18.62	88.82

Table 2: Evaluation results on the English-only Libriphrase.

Model	MSWC-IPA		FLEURS-IPA		UCLAPHONETICCORPUS		DORECO-IPA	
	Hit@1 ↑	mAP ↑	Hit@1 ↑	mAP ↑	Hit@1 ↑	mAP ↑	Hit@1 ↑	mAP ↑
CLAP-IPA-TEXT	13.51	12.7	8.48	10.22	-	-	-	-
CLAP-IPA-PHONE	79.28	68.74	86.4	87.4	-	-	-	-
CLAP-IPA-FLEURS	83.48	77.16	98.59	98.53	51.57	62.53	73.91	79.32
CLAP-IPA-VC	84.38	75.64	63.52	63.21	50.05	61.41	90.56	93.01
CLAP-IPA-tiny	82.58	76.32	98.85	98.86	51.71	62.62	95.46	96.84
CLAP-IPA-base	**82.60**	**77.31**	**99.20**	**99.27**	52.17	63.90	**96.54**	**97.77**
CLAP-IPA-small	81.98	75.35	97.61	97.98	**55.05**	**65.93**	91.46	94.41

Table 3: Evaluation results on unseen languages.

open-source models for comparison. However, we hope that our models and results will become a baseline that spur more future research in this direction.

Libriphrase To compare with existing models, we first tested on a popular English KWS dataset, Libriphrase (Shin et al., 2022), as an out-of-domain evaluation dataset, since our models were not trained on their training sets. We used **Equal Error Rate (EER)** and the **Area under Curve (AUC)** scores to compare model performance, consistent with prior studies.

Unseen languages We also evaluated all models on five unseen languages with typological diversity from FLEURS-IPA and MSWC-IPA. We isolated five language from MSWC-IPA and FLEURS-IPA, namely, Vietnamese (vie), Tamil (tam), Hausa (hau), Georgian (geo) and Odia (ori). For FLEURS-IPA, the test sets of these five languages were directly used. However, for MSWC-IPA, due to data scarcity, we pooled all training, validation, and tests of these five languages together to form a larger and more challenging benchmark. We further evaluated 95 (81 unseen) languages from the UCLA phonetic Corpus (Li et al., 2021) and 14 unseen languages from DORECO-IPA. **Hit@1** and **Mean Average Precision (mAP)** were used to measure the cross-linguistic retrieval performance of all models. To avoid duplication, we only reported results on phoneme-to-speech retrieval, as the results of speech-to-phoneme and speech-to-speech retrieval were in the same range.

Word and phoneme boundaries To evaluate the performance of forced alignment, we made use of **F1** and **R-Value**, which were used in prior studies (Räsänen et al., 2009; Kreuk et al., 2020; Zhu et al., 2022c). If the predicted boundary is within the tolerance interval of the true boundary, it is considered a hit, otherwise a miss. Since each boundary marked the onset and the offset of consecutive phones, we only evaluated the phone onsets with a tolerance of 20ms and word onsets with a tolerance of 100ms. We used TIMIT (Garofolo et al., 1993) as the English benchmark. DORECO-IPA also contains phoneme-level and word-level alignments, so we partitioned the DORECO-IPA into seen and unseen evaluation sets. Yet IPA-ALIGNER was never trained on any segmentation labels.

6 Results

In this section, we summarize the main results for KWS and forced alignment.

KWS Evaluation results in Table 2 suggests that CLAP-IPA performs on par with the state-of-

Method	TIMIT-Word		TIMIT-Phone	
	F1 ↑	R-Val ↑	F1 ↑	R-Val ↑
FAVE	-	-	58.0	64.0
MFA	-	-	63.0	68.0
Gentle	-	-	48.0	56.0
WebMAUS	-	-	**70.0**	**75.0**
W2V2-FC-20ms	-	-	48.0	56.0
W2V2-FS-20ms	-	-	48.0	55.0
ZEROSHOT				
CLAP-IPA-tiny	84.37	86.66	40.46	49.92
CLAP-IPA-base	78.61	81.73	36.16	46.59
CLAP-IPA-small	74.18	77.95	35.26	46.17
FINETUNED				
IPA-ALIGNER-tiny	**86.84**	**88.75**	57.31	63.66
IPA-ALIGNER-base	86.55	88.51	60.86	66.67
IPA-ALIGNER-small	82.33	84.76	52.54	59.53

Table 4: Evaluation of forced alignment on TIMIT. Baseline results were retrieved from Zhu et al. (2022c). The temporal resolution is 10ms for FAVE, MFA, Gentle, and WebMAUS and 20ms for the rest of the models.

the-art models on LibrisPhrase-Easy, while not trained on the Libriphrase training set. Yet CLAP-IPA failed to outperform state-of-the-art CED (Nishu et al., 2023) in LibriPhrase-Hard, suggesting that language-specific finetuning is still necessary to maximize performance. Generally speaking, phoneme-based models are more effective than text-based models.

For unseen languages, Table 3 indicates that phoneme-based models do generalize successfully to unseen languages across datasets. In contrast, the text-based model performs poorly in unseen languages, suggesting that orthographic texts are not very useful for crosslinguistic speech processing. Utterance-level retrieval appears to be much easier than word-level retrieval, a pattern quite consistent across datasets. Model size correlates with performance in seen languages but not with crosslinguistic generalizability.

Forced alignment While not trained on forced alignment explicitly, CLAP-IPA shows some capabilities for crosslinguistic forced alignment even in zero-shot predictions on both seen and unseen languages (see Table 4 and Table 5). After finetuned with the ForwardSum loss, IPA-ALIGNER can perform competitively in English with some widely used HMM-based forced aligners, even though TIMIT was not part of its training dataset. For low-resource languages, IPA-ALIGNER also achieves good performance, regardless of whether the language has been seen during training or not.

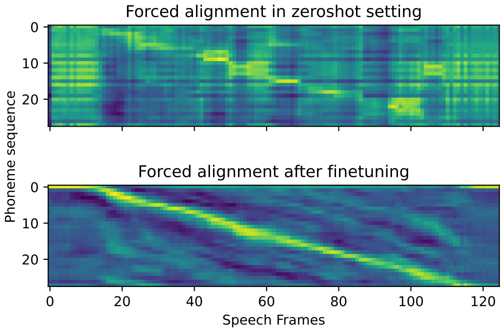

Figure 2: Illustration of forced alignment in an Evenki utterance. CLAP-IPA exhibits vague monotonic alignment without finetuning (**Top**). After finetuning, IPA-ALIGNER learns salient monotonic alignment between speech and phonemes (**Bottom**).

7 Discussions

In this section, we provide more in-depth answers to our research questions with the major findings of our experiments.

Can phoneme-based models generalize cross-linguistically? The evaluation results for CLAP-IPA and IPA-ALIGNER in Table 3 and Table 5 indicate that phoneme-based model exhibits strong generalization capabilities cross-linguistically in both KWS and forced alignment, even to unseen languages in zero-shot predictions.

Generally speaking, all CLAP-IPA models perform better on utterance-level datasets (FLEURS-IPA and DORECO-IPA) than on word-level datasets (MSWC-IPA and UCLA Phonetic Corpus), because the longer the phoneme sequence, the more likely that it is distinct in a pool of candidates. For utterance-level datasets, CLAP-IPA models achieve near-perfect scores on unseen languages (see Table 3), indicating that phonemic representations do enable cross-linguistic generalization.

Table 6 shows that the similarity assigned by CLAP-IPA-small was highly consistent with human perception. The top-ranked crosslinguistic candidates were extremely similar in articulatory features and syllable structure to the query.

For forced alignment, even the zero-shot predictions using CLAP-IPA can perform segmentation in unseen languages, especially at the word level. Interestingly, there were no significant differences between performance over seen and unseen languages. Though this result could be biased by the

Method	Seen-Word		Seen-Phone		Unseen-Word		Unseen-Phone	
	F1 ↑	R-Val ↑	F1 ↑	R-Val ↑	F1 ↑	R-Val ↑	F1 ↑	R-Val ↑
ZEROSHOT								
CLAP-IPA-tiny	67.19	72.17	33.8	44.74	68.53	73.27	35.26	45.91
CLAP-IPA-base	57.19	63.9	29.09	41.30	58.35	64.91	32.03	42.59
CLAP-IPA-small	51.48	59.43	28.59	40.85	59.94	66.18	30.24	42.43
FINETUNED								
IPA-ALIGNER-tiny	74.18	77.99	47.08	54.93	76.33	79.82	48.96	56.55
IPA-ALIGNER-base	**78.30**	**91.47**	**48.04**	**55.70**	**80.71**	**83.52**	**50.32**	**57.67**
IPA-ALIGNER-small	72.24	76.37	44.89	52.97	73.63	77.52	46.46	54.38

Table 5: Evaluation of forced alignment on DORECO-IPA. The word boundary metrics were calculated with 100ms tolerance, whereas the phone boundary was computed with 20ms tolerance.

Query	Output Type	Retrieved candidates (ranked from high to low)
étá	Most similar	ɛ̀tá, éta, lāta, étá, ìtà, aitːa, mɛ̀tá, met̪a, p͡tá, aita, ɛ̀tɛ́, atˢa, àtá, eitə
	Most dissimilar	bɔ̃t̪i, tʲuːriʃ, sorŋgi, aβuˈru, mbúruù, sumbuŋ, buluz, ʃìʒìʒ, tʃungu

Table 6: Sample ranked phonemic sequences by CLAP-IPA-small, given the speech query [étá].

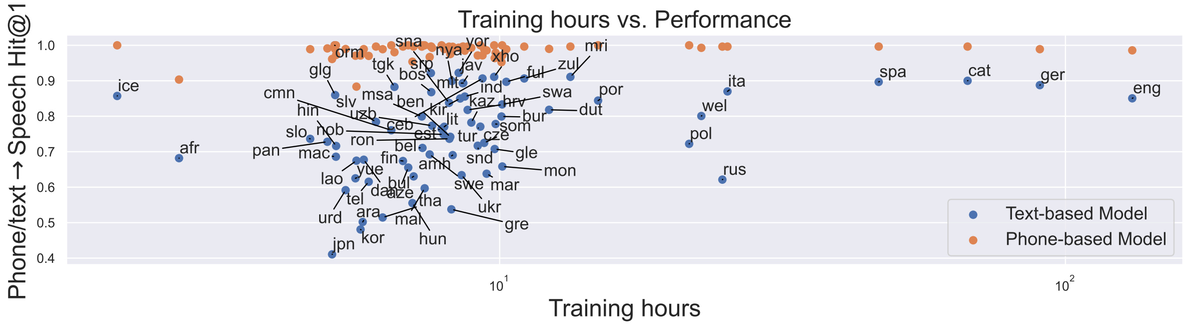

Figure 3: Correlation of model performance on individual languages with training hours by language. Languages are represented by their ISO 639-3 codes. While trained the exact same data, the phoneme-based model outperforms the text-based model in every single language, suggesting that phoneme-based modeling enables knowledge transfer across languages.

smaller number of unseen languages compared to seen languages (14 vs. 30), it still suggests that IPA-ALIGNER can perform crosslinguistic forced alignment without much adaptation. Finetuning the IPA-ALIGNER brings continued improvement over the zero-shot scenarios (see Fig 2).

Does the phoneme-based model generalize better cross-linguistically than the text-based model? Text-based models are struggling to generalize to unseen languages, as these unseen languages have their distinct writing systems (e.g, Vietnamese and Tamil) that are not seen in training languages. Comparison between the text-based and phoneme-based models in Table 2 and Table 3 clearly shows that it is the use of phonemes as modeling units that brings strong crosslinguistic generalizability, since they can represent all languages using the same set of symbols.

Do the training hours of individual languages predict the performance of multilingual models? The number of training hours for individual languages does not predict the performance of language in phoneme-based models. All languages benefit from the multilingual knowledge transfer in phoneme-based modeling.

It has been reported that there is a strong correlation between text-based multilingual ASR performance in individual languages and their training hours (Radford et al., 2023; Rouditchenko et al., 2023). We also confirm that, for text-based models, there is a moderate correlation between Hit@1 and the number of training hours (Spearman's $\rho : 0.42$;

$p \leq 0.0002$). However, this correlation was not significant for the phoneme-based model (Spearman's ρ : 0.14; $p = 0.22$). In Figure 3, the phoneme-based model outperforms the text-based model in every language by a large margin, especially for languages with less training data.

Since the orthography varies across languages and is usually not an accurate reflection of pronunciation, many low-resource languages are not reaping the full benefits of large-scale multilingual data in this cross-modal task in text-based models. Close inspection shows that the text-based model generalizes well to Hausa (Latin alphabet) but significantly underperforms in languages with non-Latin alphabet, such as Tamil, Vietnamese, Japanese, Arabic, and Cantonese.

In contrast, the phoneme-based model achieves near-perfect performance in retrieval in almost all seen and unseen languages (see Figure 3), making them extremely useful in low-resource and zero-resource scenarios. The efficiency of IPA representations in multilingual settings has also been observed in ASR (Feng et al., 2023).

Does multilingual models always hold advantages over monolingual models? At least in the current study, multilingual models might not hold an apparent advantage over well-engineered monolingual models in high-resource languages. As shown in Table 2 and Table 4, compared to other state-of-the-art KWS and forced alignment models, CLAP-IPA and IPA-ALIGNER was not able to outperform well-engineered monolingual models. Our multilingual models have not been trained on the training set of LibriPhrase or TIMIT, so some of the performance gaps might be caused by domain mismatch. Even with zero adaptations, multilingual models achieve close performance to monolingual models, suggesting that our approach is promising and may reach better results if scaled up.

Should we scale up the number of languages or number of training hours? We compared CLAP-IPA only trained on VoxCommunis (Ahn and Chodroff, 2022) or FLEURS-IPA. VoxCommunis has almost twice as many hours as FLEURS-IPA with roughly half of the languages. In Table 2 and 3, CLAP-IPA-VC trained on more hours of speech generally has similar performance as CLAP-IPA-FLEURS trained on a subset of the IPAPACK across metrics, which suggests that creating high-quality data is effective in achieving good performance. But this finding also suggests that we can achieve good crosslinguistic generalizability with fewer languages but longer hours using phoneme modeling. Given the empirical data distributions in real-life settings, scaling up training hours in a dozen of languages is much easier than scaling up the number of languages. The practical implication is that we might be able to build multilingual speech processing systems for many low-resource or zero-resource languages with large-scale data in a dozen relatively high-resource languages.

Is it feasible to scale up the creation of good-quality phonemic transcriptions in world languages? Despite our attempt, there are still multiple challenges for creating phonemic transcriptions. During our dataset construction, we were unable to process many languages due to the lack of pronunciation dictionaries, text transcriptions, or relevant NLP tools, especially the lack of good word segmentation tools for some East/Southeast Asian languages like Khmer. While available large-scale speech corpora nowadays encompass more than 1000 languages (Salesky et al., 2020; Pratap et al., 2023), textual or phonemic labels cannot be easily obtained for most of them, limiting their usage in many research applications.

Even for high-resource languages, preprocessing multilingual texts and normalizing the Unicode encodings for IPA symbols usually take tremendous effort, not to mention verifying these phonemic transcriptions for audio recordings. It remains unclear how biases or noises in G2P predictions will propagate to downstream multilingual tasks. Our endeavor marks a small step in creating good-quality phonemic transcriptions for more languages. However, there is still much work to be done to include a broader array of languages worldwide and to improve the quality of transcriptions.

8 Conclusions

With the carefully curated IPAPACK, we show that using IPA symbols as modeling units can effectively enable CLAP-IPA and IPA-ALIGNER to generalize to unseen languages, highlighting the benefits of incorporating linguistic knowledge into deep learning methods. We believe that the IPAPACK has great potential to benefit more tasks in multilingual speech processing, such as multilingual phoneme recognition, speech synthesis, and documenting endangered languages. In the future, we will continue to expand our dataset and models to include more diverse languages.

9 Ethical statement

Data Governance We adhered strictly to ethical practices in curating our datasets. The original FLEURS (Conneau et al., 2023), MSWC (Mazumder et al., 2021b), DoReCo (Paschen et al., 2020) and VoxCommunis (Ahn and Chodroff, 2022) corpora are distributed under the Creative Commons licenses. Therefore, we are permitted to re-process and re-distribute the original dataset with proper attributions. Some languages in the DoReCo corpus are under a Creative Commons Non-Commercial license. We reserved these languages to the test set in our corpora, such that our models have not been trained on data under commercially restrictive licenses. As required, we have also cited every individual language from the DoReCo Corpus in Table 11.

Potential Impact We believe that our dataset and models will contribute to the endeavor of building fair and inclusive speech processing systems for all languages and facilitating the documentation of endangered languages. However, we are aware that multilingual keyword-spotting technology could potentially be misused as surveillance tools for monitoring speech recordings in more languages, posing risks to users.

10 Limitations

Our study is still limited in several aspects. First, while we tried our best to inspect a subset of our dataset, it was impossible for us to examine all datasets in great detail. As a result, the constructed dataset might still be flawed in terms of audio quality and transcription quality (and many unicode errors). Secondly, the proposed models are still not optimized in terms of computational efficiency. Since most KWS applications are running on mobile devices with limited computational power, the proposed models still have too many model parameters to run efficiently on mobile devices. Moreover, speech sequences are usually much longer than text sequences. Self-attention with quadratic complexity might not be the most suitable architecture for processing speech. More efforts are needed to make such multilingual models efficient.

Thirdly, the number of languages studied in our paper is still limited and might be biased towards languages that are relatively high-resource. They are not representative of the global language landscape. There are many more low-resource or endangered languages we are not able to include due to the lack of various resources. To promote linguistic inclusion and fairness, we will continue to improve the language diversity of our research in the future.

Acknowledgements

This research was enabled in part through the computational resources and services provided by Advanced Research Computing at the University of British Columbia and in part through the support provided by the Digital Research Alliance of Canada. This study also benefits from the cloud computing credits awarded by Microsoft Azure through a pilot program with UBC Advanced Research Computing.

The authors would like to thank the four anonymous reviewers as well as the area chairs for their thoughtful comments and discussions, which helped improve this article considerably. We thank Emily P. Ahn and Eleanor Chodroff for creating and releasing the VoxCommunis, which inspired our data creation process. Finally, we acknowledge that this work is impossible without the pioneering efforts of many language researchers who collected and shared speech corpora across world languages.

References

Emily P. Ahn and Eleanor Chodroff. 2022. VoxCommunis: A corpus for cross-linguistic phonetic analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5286–5294, Marseille, France. European Language Resources Association.

Mathieu Avanzi, Marie-José Béguelin, Gilles Corminboeuf, Federica Diémoz, and Laure Anne Johnsen. 2022. French (Swiss) DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Jocelyn Aznar. 2022. Nisvai DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

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*.

Rohan Badlani, Adrian Łańcucki, Kevin J Shih, Rafael Valle, Wei Ping, and Bryan Catanzaro. 2022. One tts alignment to rule them all. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6092–6096. IEEE.

Axel Berg, Mark O'Connor, and Miguel Tairum Cruz. 2021. Keyword Transformer: A Self-Attention Model for Keyword Spotting. In *Proc. Interspeech 2021*, pages 4249–4253.

Natalia Bogomolova, Dmitry Ganenkov, and Nils Norman Schiborr. 2022. Tabasaran DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Niclas Burenhult. 2022. Jahai DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Guoguo Chen, Carolina Parada, and Georg Heigold. 2014. Small-footprint keyword spotting using deep neural networks. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4087–4091. IEEE.

Yi-Chang Chen, Yu-Chuan Steven, Yen-Cheng Chang, and Yi-Ren Yeh. 2022. g2pW: A Conditional Weighted Softmax BERT for Polyphone Disambiguation in Mandarin. In *Proc. Interspeech 2022*, pages 1926–1930.

Alexander Yao Cobbinah. 2022. Baïnounk Gubëeher DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

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*.

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

Andrew Cowell. 2022. Arapaho DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul-Ambroise Duquenne, Hongyu Gong, and Holger Schwenk. 2021. Multimodal and multilingual embeddings for large-scale speech mining. *Advances in Neural Information Processing Systems*, 34:15748–15761.

Chris Lasse Däbritz, Nina Kudryakova, Eugénie Stapert, and Alexandre Arkhipov. 2022. Dolgan DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Christian Döhler. 2022. Komnzo DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Siyuan Feng, Ming Tu, Rui Xia, Chuanzeng Huang, and Yuxuan Wang. 2023. Language-Universal Phonetic Representation in Multilingual Speech Pretraining for Low-Resource Speech Recognition. In *Proc. INTERSPEECH 2023*, pages 1384–1388.

Diana Forker and Nils Norman Schiborr. 2022. Sanzhi Dargwa DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Michael Franjieh. 2022. Fanbyak DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Alexandro Garcia-Laguia. 2022. Northern Alta DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. 1993. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403.

Bryan Gick, Ian Wilson, and Donald Derrick. 2013. *Articulatory phonetics*. John Wiley & Sons.

Jost Gippert. 2022. Svan DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Kevin Glocker, Aaricia Herygers, and Munir Georges. 2023. Allophant: Cross-lingual Phoneme Recognition with Articulatory Attributes. In *Proc. INTERSPEECH 2023*, pages 2258–2262.

Matthew Gordon and Timo Roettger. 2017. Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard*, 3(1):20170007.

Matthew K Gordon. 2016. *Phonological typology*, volume 1. Oxford University Press.

Richard Griscom. 2022. Asimjeeg Datooga DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Valentin Gusev, Tiina Klooster, Beáta Wagner-Nagy, and Alexandre Arkhipov. 2022. Kamas DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Tom Güldemann, Martina Ernszt, Sven Siegmund, and Alena Witzlack-Makarevich. 2022. N||ng DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Geoff Haig, Maria Vollmer, and Hanna Thiele. 2022. Northern Kurdish (Kurmanji) DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Andrew Harvey. 2022. Gorwaa DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Berry Heselwood. 2013. *Phonetic Transcription in Theory and Practice*. Edinburgh University Press.

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

Olga Kazakevich and Elena Klyachko. 2022. Evenki DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Matthew C. Kelley and Benjamin V. Tucker. 2018. A Comparison of Input Types to a Deep Neural Network-based Forced Aligner. In *Proc. Interspeech 2018*, pages 1205–1209.

Sameer Khurana, Antoine Laurent, and James Glass. 2022. Samu-xlsr: Semantically-aligned multimodal utterance-level cross-lingual speech representation. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1493–1504.

Soung-U Kim. 2022. Jejuan DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Thomas Kisler, Florian Schiel, and Han Sloetjes. 2012. Signal processing via web services: the use case webmaus. In *Digital Humanities Conference 2012*.

Felix Kreuk, Joseph Keshet, and Yossi Adi. 2020. Self-Supervised Contrastive Learning for Unsupervised Phoneme Segmentation. In *Proc. Interspeech 2020*, pages 3700–3704.

Manfred Krifka. 2022. Daakie DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

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. 1990. The revised international phonetic alphabet. *Language*, 66(3):550–552.

Peter Ladefoged and Morris Halle. 1988. Some major features of the international phonetic alphabet. *Language*, 64(3):577–582.

Yong-Hyeok Lee and Namhyun Cho. 2023. PhonMatchNet: Phoneme-Guided Zero-Shot Keyword Spotting for User-Defined Keywords. In *Proc. INTERSPEECH 2023*, pages 3964–3968.

Lei Lei, Guoshun Yuan, Hongjiang Yu, Dewei Kong, and Yuefeng He. 2023. Multilingual customized keyword spotting using similar-pair contrastive learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

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.

Mark Mazumder, Colby Banbury, Josh Meyer, Pete Warden, and Vijay Janapa Reddi. 2021a. Few-Shot Keyword Spotting in Any Language. In *Proc. Interspeech 2021*, pages 4214–4218.

Mark Mazumder, Sharad Chitlangia, Colby Banbury, Yiping Kang, Juan Manuel Ciro, Keith Achorn, Daniel Galvez, Mark Sabini, Peter Mattson, David Kanter, et al. 2021b. Multilingual spoken words corpus. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Paul McCann. 2020. fugashi, a tool for tokenizing Japanese in python. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 44–51, Online. Association for Computational Linguistics.

Alexis Michaud. 2022. Yongning Na DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

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)*.

Ulrike Mosel. 2022. Teop DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Kumari Nishu, Minsik Cho, Paul Dixon, and Devang Naik. 2023. Flexible keyword spotting based on homogeneous audio-text embedding. *arXiv preprint arXiv:2308.06472*.

Kari A Noriy, Xiaosong Yang, Marcin Budka, and Jian Jun Zhang. 2023. Clara: Multilingual contrastive learning for audio representation acquisition. *arXiv preprint arXiv:2310.11830*.

Carmel O'Shannessy. 2022a. Light Warlpiri DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Carmel O'Shannessy. 2022b. Warlpiri DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Pavel Ozerov. 2022. Anal DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

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. *Interspeech 2019*.

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.

Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, and Pattarawat Chormai. 2016. PyThaiNLP: Thai Natural Language Processing in Python.

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.

Maïa Ponsonnet. 2022. Dalabon DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

Juan Diego Quesada, Stavros Skopeteas, Carolina Pasamonik, Carolin Brokmann, and Florian Fischer. 2022. Cabécar DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

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.

Sabine Reiter. 2022. Cashinahua DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Paul M Reuter, Christian Rollwage, and Bernd T Meyer. 2023. Multilingual query-by-example keyword spotting with metric learning and phoneme-to-embedding mapping. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Sonja Riesberg. 2022. Yali (Apahapsili) DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Hiram Ring. 2022. Pnar DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Françoise Rose. 2022. Mojeño Trinitario DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Ingrid Rosenfelder, Josef Fruehwald, Keelan Evanini, and Jiahong Yuan. 2011. Fave (forced alignment and vowel extraction) program suite. *URL http://fave. ling. upenn. edu*.

Andrew Rouditchenko, Sameer Khurana, Samuel Thomas, Rogerio Feris, Leonid Karlinsky, Hilde Kuehne, David Harwath, Brian Kingsbury, and James Glass. 2023. Comparison of Multilingual Self-Supervised and Weakly-Supervised Speech Pre-Training for Adaptation to Unseen Languages. In *Proc. INTERSPEECH 2023*, pages 2268–2272.

Oleg Rybakov, Natasha Kononenko, Niranjan Subrahmanya, Mirkó Visontai, and Stella Laurenzo. 2020. Streaming Keyword Spotting on Mobile Devices. In *Proc. Interspeech 2020*, pages 2277–2281.

Okko Johannes Räsänen, Unto Kalervo Laine, and Toomas Altosaar. 2009. An improved speech segmentation quality measure: the r-value. In *Proc. Interspeech 2009*, pages 1851–1854.

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.

Stefan Schnell. 2022. Vera'a DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Kilian Schulze-Forster, Clement SJ Doire, Gaël Richard, and Roland Badeau. 2020. Joint phoneme alignment and text-informed speech separation on highly corrupted speech. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7274–7278. IEEE.

Frank Seifart. 2022a. Bora DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Frank Seifart. 2022b. Resígaro DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Kevin J Shih, Rafael Valle, Rohan Badlani, Adrian Lancucki, Wei Ping, and Bryan Catanzaro. 2021. Rad-tts: Parallel flow-based tts with robust alignment learning and diverse synthesis. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*.

Hyeon-Kyeong Shin, Hyewon Han, Doyeon Kim, Soo-Whan Chung, and Hong-Goo Kang. 2022. Learning Audio-Text Agreement for Open-vocabulary Keyword Spotting. In *Proc. Interspeech 2022*, pages 1871–1875.

Stavros Skopeteas, Violeta Moisidi, Nutsa Tsetereli, Johanna Lorenz, and Stefanie Schröter. 2022. Urum DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Kazuyo Tanaka, Yoshiaki Itoh, Hiroaki Kojima, and Nahoko Fujimura. 2001. Speech data retrieval system constructed on a universal phonetic code domain. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU'01.*, pages 323–326. IEEE.

Raphael Tang and Jimmy Lin. 2018. Deep residual learning for small-footprint keyword spotting. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5484–5488. IEEE.

Amos Teo. 2022. Sümi DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Yann Teytaut, Baptiste Bouvier, and Axel Roebel. 2022. A study on constraining connectionist temporal classification for temporal audio alignment. In *Interspeech 2022*, pages 5015–5019. ISCA.

Yann Teytaut and Axel Roebel. 2021. Phoneme-to-audio alignment with recurrent neural networks for speaking and singing voice. In *Proceedings of Interspeech 2021*, pages 61–65. International Speech Communication Association; ISCA.

Nick Thieberger. 2022. Nafsan (South Efate) DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Martine Vanhove. 2022. Beja DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Alexandra Vydrina. 2022. Kakabe DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Claudia Wegener. 2022. Savosavo DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

John C Wells. 1995. Computer-coding the IPA: a proposed extension of SAMPA.

Søren Wichmann. 2022. Texistepec Popoluca DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Alena Witzlack-Makarevich, Saudah Namyalo, Anatol Kiriggwajjo, and Zarina Molochieva. 2022. Ruuli DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023.

Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2022. Simple and Effective Zero-shot Cross-lingual Phoneme Recognition. In *Proc. Interspeech 2022*, pages 2113–2117.

Xianming Xu and Bibo Bai. 2022. Sadu DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*.

Jian Zhu, Zuoyu Tian, Yadong Liu, Cong Zhang, and Chia-Wen Lo. 2022a. Bootstrapping meaning through listening: Unsupervised learning of spoken sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1134–1154, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Jian Zhu, Cong Zhang, and David Jurgens. 2022c. Phone-to-audio alignment without text: A semi-supervised approach. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8167–8171. IEEE.

A Dataset statistics

Table 9, 12 , 10, and 11 provide tabulated summaries of the detailed statistics of our curated datasets.

B Training hyperparameters

For pre-training, we trained three variants of BERT from scratch using only phonemic transcriptions. We adopted the AdamW optimizer with an initialized learning rate of $1e-4$ and cosine scheduling with a warm-up step of 1000. All models were trained for 60k iterations before stopping. All training processes were completed on a single V100 GPU of 32 GB.

All hyperparameters for CLAP-IPA models were listed in Table 7. By default, all models were trained on a single V100 with 32GB of memory. The training time for CLAP-IPA in 100k steps ranged from 17 hours for CLAP-IPA-tiny to 41 hours for CLAP-IPA-small.

All hyperparameters for IPA-ALIGNER models were listed in Table 8. All models were trained on a single V100 with 32GB of memory. The training time for IPA-ALIGNER before early stopping ranged from 5 hours for CLAP-IPA-tiny to 12 hours for CLAP-IPA-small.

Hyperparameters	CLAP-IPA-tiny	CLAP-IPA-base	CLAP-IPA-small
Hidden dimensions	384	512	768
Num. Layers	4	6	12
Num. Att. Heads	6	8	12
Intermediate size	1536	2048	3072
Parameters	16M	28.5M	96.2M
Initial learning rate	1e-4		
Scheduler	Cosine Scheduler		
Warm-up steps	500		
Total training steps	100k		
FLEURS-IPA batch size	64	64	32
MSWC-IPA batch size	512	512	256
DORECO-IPA batch size	64	64	32
VoxCommunis batch size	64	64	32
Gradient checkpointing	True		
Mixed Precision	float16		
Max. Gradient Norm for Gradient Clipping	10		

Table 7: Hyperparameters for training CLAP-IPA models.

Hyperparameters	CLAP-IPA-tiny	CLAP-IPA-base	CLAP-IPA-small
Hidden dimensions	384	512	768
Num. Layers	4	6	12
Num. Att. Heads	6	8	12
Intermediate size	1536	2048	3072
Parameters	16M	28.5M	96.2M
Initial learning rate	1e-5		
Scheduler	Cosine Scheduler		
Warm-up steps	100		
Maximum training steps	10k		
batch size	128		
Gradient checkpointing	True		
Mixed Precision	float16		
Max. Gradient Norm for Gradient Clipping	10		
Early stopping	True		
Stopping Criteria	the highest F1 on the TIMIT training set		

Table 8: Hyperparameters for training IPA-ALIGNER models.

Language	ISO 639-3	Family	Train (hrs)	Dev (hrs)	Test (hrs)	Avg.Phones	Avg.Dur. (s)
Arabic	ara	Indo-European	0.79	0.11	0.11	6.57 (1.58)	1 (0)
Catalan	cat	Indo-European	61.38	8.12	8.07	7.51 (2.34)	1 (0)
Czech	cze	Indo-European	3.01	0.41	0.39	6.2 (1.92)	1 (0)
Dutch	dut	Indo-European	6.42	0.84	0.84	7.2 (2.79)	1 (0)
English	eng	Indo-European	125.61	16.29	16.52	6.84 (2.34)	1 (0)
Esperanto	epo	Constructed	8.48	1.13	1.12	6.95 (1.97)	1 (0)
Estonian	est	Uralic	2.51	0.34	0.33	6.35 (2.13)	1 (0)
French	fra	Indo-European	62.69	8.31	8.31	5.8 (2.03)	1 (0)
German	ger	Indo-European	83.23	10.99	10.96	8.81 (3.41)	1 (0)
Irish	gle	Indo-European	0.48	0.07	0.07	4.17 (1.45)	1 (0)
Greek	gre	Indo-European	0.71	0.1	0.1	5.71 (2.09)	1 (0)
Interlingua	ina	Constructed	0.53	0.06	0.05	5.8 (1.84)	1 (0)
Indonesian	ind	Austronesian	1.74	0.25	0.25	6.04 (1.88)	1 (0)
Italian	ita	Indo-European	18.42	2.46	2.43	7.35 (2.45)	1 (0)
Kyrgyz	kir	Turkic	1.52	0.23	0.21	6.89 (2.43)	1 (0)
Lithuanian	lit	Indo-European	0.7	0.04	0.06	6.88 (2.22)	1 (0)
Maltese	mlt	Afro-Asiatic	1.12	0.15	0.16	6.04 (2.75)	1 (0)
Mongolian	mon	Mongolic	1.48	0.2	0.21	5.5 (1.73)	1 (0)
Polish	pol	Indo-European	14.39	1.93	1.94	6.9 (2.09)	1 (0)
Portuguese	por	Indo-European	7.16	0.95	0.95	6.52 (2.07)	1 (0)
Romanian	ron	Indo-European	0.5	0.1	0.08	6.43 (2.31)	1 (0)
Russian	rus	Indo-European	18.48	2.46	2.44	8.48 (2.88)	1 (0)
Slovak	slo	Indo-European	0.08	0.01	0.01	6.28 (2.14)	1 (0)
Slovenian	slv	Indo-European	0.27	0.05	0.05	5.16 (1.56)	1 (0)
Spanish	spa	Indo-European	40.04	5.35	5.32	7.64 (2.49)	1 (0)
Swedish	swe	Indo-European	1.18	0.16	0.16	5.36 (1.93)	1 (0)
Tatar	tat	Turkic	3.76	0.5	0.48	6.18 (1.84)	1 (0)
Turkish	tur	Turkic	2.82	0.38	0.39	6.95 (2.27)	1 (0)
Ukrainian	ukr	Indo-European	1.87	0.25	0.26	6.76 (2.3)	1 (0)
Welsh	wel	Indo-European	13.6	1.8	1.81	5.76 (1.96)	1 (0)
Mandarin	cmn	Sino-Tibetan	0.4	0.05	0.05	8.68 (2.26)	1 (0)

Table 9: Statistics of languages in MSWC-IPA. All samples are padded to be clips of 1 second. (Avg.Phones: average number of phonemes in each word; Avg.Dur.: average duration of each clip).

Language	ISO 639-3	Family	Train (hrs)	Avg. Dur (s)	Avg. Phones
Abkhaz	abk	Northwest Caucasian	0.62	7.33	51.93
Bashkir	bak	Turkic	137.79	4.35	35.78
Belarusian	bel	Indo-European	132.21	5.48	49.15
Bulgarian	bul	Indo-European	3.5	5.05	47.74
Catalan	cat	Indo-European	2.08	5.39	44.35
Czech	ces	Indo-European	16.51	4.75	44
Chuvash	chv	Turkic	0.37	4.2	36.97
Greek	ell	Indo-European	1.57	3.99	29.13
Basque	eus	Language isolate	12.66	5.2	47.36
Guarani	grn	Tupian	1.81	3.97	26.91
Hausa	hau	Afro-Asiatic	1.71	4.27	32.1
Hindi	hin	Indo-European	2.73	3.75	33.69
Sorbian (Upper Sorbian)	hsb	Indo-European	1.48	6.61	55.01
Hungarian	hun	Uralic	25.06	4.76	37.62
Indonesian	ind	Austronesian	5.09	5.69	53.31
Italian	ita	Indo-European	192.69	5.24	49.13
Georgian	kat	Kartvelian	1.62	5.77	53.7
Kazakh	kaz	Turkic	0.29	4.93	33.95
Kurmanji (Kurdish)	kmr	Indo-European	2.83	4.47	28.16
Kyrgyz	kir	Turkic	2.25	4.67	43.42
Marathi	mar	Indo-European	3.66	5.97	52.99
Maltese	ml	Afro-Asiatic	2.41	4.48	36.88
Erzya	myv	Uralic	1.97	5.73	46.41
Dutch	nld	Indo-European	34.94	4.4	47.47
Punjabi	pan	Indo-European	0.96	5.29	26.53
Polish	pol	Indo-European	14.26	5.21	46.58
Portuguese	por	Indo-European	12.31	4.33	32.45
Romanian	ron	Indo-European	4.99	4.01	35.98
Russian	rus	Indo-European	24.41	5.44	56.61
Swedish	swe	Indo-European	5.84	3.84	31.24
Swahili	swa	Niger-Congo	52.8	5.44	47.43
Tamil	tam	Dravidian	61.39	6.57	55.43
Thai	tha	Kra-Dai	16.71	3.91	26.58
Turkish	tur	Turkic	0.98	3.19	30.43
Tatar	tat	Turkic	10.09	3.8	31.5
Uyghur	uig	Turkic	2.43	5.85	49.21
Ukrainian	ukr	Indo-European	4.22	4.67	39.54
Vietnamese	vie	Austroasiatic	4.6	4.53	25.54

Table 10: Detailed statistics of a selected subset of VoxCommunis (Ahn and Chodroff, 2022). (Avg.Phones: average number of phonemes in each word; Avg.Dur.: average duration of each clip).

Language	ISO 693-3	Avg. Dur (s)	Total duration (hrs)	Avg. Phones	Family	Split	Citation
Komnzo	tci	2.59	0.27	29.99	Yam	train	(Döhler, 2022)
Vera'a	vra	3.55	0.57	43.03	Austronesian	train	(Schnell, 2022)
Sanzhi Dargwa	na	4.85	0.17	44.82	Nakh-Daghestanian	train	(Forker and Schiborr, 2022)
Urum	uum	4.63	0.37	45.75	Turkic	test	(Skopeteas et al., 2022)
Beja	bej	2.32	0.36	24.97	Afro-Asiatic	test	(Vanhove, 2022)
Light Warlpiri	na	3.47	0.47	32.75	Mixed Language	train	(O'Shannessy, 2022a)
Kamas	xas	3.60	0.84	24.71	Uralic	train	(Gusev et al., 2022)
Nafsan (South Efate)	erk	6.10	0.36	50.83	Austronesian	test	(Thieberger, 2022)
Tabasaran	tab	4.16	0.21	42.31	Nakh-Daghestanian	train	(Bogomolova et al., 2022)
Savosavo	svs	5.17	0.82	49.35	Isolate	train	(Wegener, 2022)
Sümi	nsm	2.74	0.14	32.59	Sino-Tibetan	train	(Teo, 2022)
French (Swiss)	fra	2.75	0.31	32.61	Indo-European	test	(Avanzi et al., 2022)
Northern Alta	aqn	2.78	1.04	25.94	Austronesian	train	(Garcia-Laguia, 2022)
Jejuan	jje	2.59	0.03	24.43	Koreanic	train	(Kim, 2022)
Jahai	jhi	3.61	0.45	32.74	Austroasiatic	test	(Burenhult, 2022)
Nisvai	none	3.11	0.56	42.22	Austronesian	test	(Aznar, 2022)
Warlpiri	wbp	3.64	0.94	30.84	Pama-Nyungan	test	(O'Shannessy, 2022b)
Fanbyak	fnb	2.81	0.22	27.29	Austronesian	train	(Franjieh, 2022)
Bora	boa	4.40	0.34	41.49	Boran	train	(Seifart, 2022a)
Yongning Na	nru	4.23	0.30	33.15	Sino-Tibetan	train	(Michaud, 2022)
Dalabon	ngk	2.46	0.08	23.46	Gunwinyguan	train	(Ponsonnet, 2022)
Sadu	na	2.75	0.15	22.78	Sino-Tibetan	train	(Xu and Bai, 2022)
Teop	tio	2.96	0.65	30.62	Austronesian	train	(Mosel, 2022)
Cashinahua	cbs	3.58	0.73	33.55	Pano-Tacanan	train	(Reiter, 2022)
Dolgan	dlg	4.24	0.69	43.55	Turkic	test	(Däbritz et al., 2022)
Anal	anm	3.02	0.37	26.43	Sino-Tibetan	train	(Ozerov, 2022)
Baïnounk Gubëeher	bab	3.13	0.40	30.95	Atlantic-Congo	train	(Cobbinah, 2022)
Texistepec Popoluca	poq	2.65	0.08	28.50	Mixe-Zoque	train	(Wichmann, 2022)
Daakie	ptv	3.22	0.22	34.87	Austronesian	train	(Krifka, 2022)
Ning	ngh	2.67	0.12	22.96	Tuu	train	(Güldemann et al., 2022)
Ruuli	ruc	3.13	0.32	34.99	Atlantic-Congo	train	(Witzlack-Makarevich et al., 2022)
Cabécar	cjp	3.61	0.38	39.62	Chibchan	test	(Quesada et al., 2022)
Evenki	evn	3.89	0.66	31.71	Tungusic	train	(Kazakevich and Klyachko, 2022)
Arapaho	arp	3.99	0.87	32.95	Algic	train	(Cowell, 2022)
Svan	sva	4.77	0.56	47.85	Kartvelian	train	(Gippert, 2022)
Resígaro	rgr	5.45	1.27	33.31	Arawakan	train	(Seifart, 2022b)
Yali (Apahapsili)	na	2.38	0.04	32.34	Nuclear Trans New Guinea	test	(Riesberg, 2022)
Asimjeeg Datooga	na	2.81	0.28	28.30	Nilotic	train	(Griscom, 2022)
Northern Kurdish (Kurmanji)	kmr	4.39	0.54	50.76	Indo-European	test	(Haig et al., 2022)

Gorwaa	gow	2.95	0.28	31.29	Afro-Asiatic	train	(Harvey, 2022)
Pnar	pbv	8.28	0.29	72.74	Austroasiatic	test	(Ring, 2022)
Kakabe	kke	4.14	0.59	33.07	Mande	train	(Vydrina, 2022)
Mojeño Trinitario	trn	5.66	0.65	48.42	Arawakan	train	(Rose, 2022)

Table 11: Detailed statistics of DORECO-IPA. (Avg. Phones: average number of phonemes in each word; Avg.Dur.: average duration of each clip).

Language	ISO 639-3	Family	Train (hrs)	Dev (hrs)	Test (hrs)	Avg. Dur (s)	Avg. Phones
Afrikaans	afr	Indo-European	2.71	0.48	0.66	11.95	4.69
Amharic	amh	Afro-Asiatic	8.26	0.57	1.28	11.91	7.25
Arabic	ara	Afro-Asiatic	4.93	0.75	1.12	10.2	6.63
Azerbaijani	aze	Turkic	6.89	1.1	2.42	12.27	6.33
Belarusian	bel	Indo-European	7.3	1.37	3.11	13.87	6.08
Bulgarian	bul	Indo-European	7.05	0.85	1.44	10.65	5.47
Bengali	ben	Indo-European	8.18	1.21	2.75	12.67	5.84
Bosnian	bos	Indo-European	7.57	1.1	2.47	11.4	5.42
Catalan	cat	Indo-European	5.77	1.09	2.43	11.34	4.64
Cebuano	ceb	Austronesian	9.33	0.72	1.77	13.26	4.54
Mandarin Chinese	cmn	Sino-Tibetan	6.04	0.87	2	10.37	3.81
Czech	cze	Indo-European	6.38	0.82	1.91	10.76	5.58
Welsh	wel	Indo-European	9.12	1.49	3.32	12.98	4.27
Danish	dan	Indo-European	5.75	0.99	2.26	10.69	4.4
German	ger	Indo-European	6.88	1.06	2.46	11.16	5.61
Greek	gre	Indo-European	7.51	0.64	1.47	10.69	5.14
English	eng	Indo-European	5.64	0.88	1.39	9.79	4.41
Spanish	spa	Indo-European	6.73	1.17	2.45	11.24	4.87
Estonian	est	Uralic	5.38	1.02	2.37	10.57	6.35
Fula	ful	Niger-Congo	10.27	0.84	2.12	14.35	4.16
Finnish	fin	Uralic	6.75	1.18	2.58	11.61	7.04
Irish	gle	Indo-European	9.31	1.24	2.76	14.54	3.68
Galician	glg	Indo-European	5.12	0.89	2.06	10.31	5.01
Hausa	hau	Afro-Asiatic	10.09	1.25	2.47	15.21	4.36
Hindi	hin	Indo-European	5.14	0.63	1.11	11.01	4.08
Croatian	hrv	Indo-European	8.78	0.85	1.98	11.14	5.43
Hungarian	hun	Uralic	7.01	1.14	2.45	10.85	5.76
Indonesian	ind	Austronesian	6.94	0.97	1.89	12.18	5.79
Icelandic	ice	Indo-European	2.11	0.1	0.14	10.8	5.53
Italian	ita	Indo-European	6.86	1.31	2.8	11.52	4.97
Japanese	jpn	Japonic	5.06	0.67	1.52	11.63	3.63
Javanese	jav	Austronesian	8.6	0.94	2.22	12.98	5.47
Georgian	geo	Kartvelian	3.87	0.99	2.37	11.31	7.14
Kazakh	kaz	Turkic	8.91	1.29	3.02	13.55	6.78
Korean	kor	Koreanic	5.68	0.57	1.03	12.14	7.17
Kyrgyz	kir	Turkic	6.99	1.1	2.52	11.45	6.83
Lao	lao	Kra-Dai	5.58	0.47	1.09	13.41	21.33
Lithuanian	lit	Indo-European	7.28	0.97	2.32	10.96	6.33
Maori	mri	Austronesian	13.34	1.86	4.53	19.34	3.48
Macedonian	mac	Indo-European	5.14	1.05	2.47	10.5	5.35
Malayalam	mal	Indo-European	7.37	1.36	2.86	12.28	10.23
Mongolian	mon	Mongolic	8.63	0.97	2.21	12.19	5.52
Marathi	mar	Indo-European	9.48	1.23	3.04	12.96	6
Malay	msa	Austronesian	7.28	0.79	1.82	11.8	5.99
Maltese	mlt	Afro-Asiatic	7.5	1.24	2.81	12.31	4.68
Burmese	bur	Sino-Tibetan	10.07	1.49	3.25	14.56	12.17
Norwegian	nob	Indo-European	7.96	0.43	0.93	12.06	4.54
Dutch	dut	Indo-European	5.81	0.38	0.77	9.18	4.89
Nyanja	nya	Niger-Congo	8.23	1.2	2.77	14.53	5.99
Oromo	orm	Afro-Asiatic	5.11	0.05	0.13	13.46	5.41
Oriya	ori	Indo-European	2.42	1	2.25	11.33	6.5
Punjabi	pan	Indo-European	4.96	0.63	1.48	11.49	4.07
Polish	pol	Indo-European	7.23	0.73	1.63	10.71	5.66
Portuguese	por	Indo-European	7.77	1.06	2.5	12.49	4.7
Romanian	ron	Indo-European	7.65	0.88	1.95	11.46	5.31
Russian	rus	Indo-European	6.28	0.92	1.94	10.97	6.24
Sindhi	snd	Indo-European	9.15	1.1	2.55	12.15	4.41
Slovak	slo	Indo-European	4.55	0.92	2.1	10.8	5.58
Slovenian	slv	Indo-European	5.78	0.74	1.78	10.2	5.43
Shona	sna	Niger-Congo	7.56	1.27	3.03	14.12	6.88
Somali	som	Afro-Asiatic	9.84	1.26	3.03	14.04	4.77
Serbian	srp	Indo-European	8.14	0.7	1.66	12.05	5.25
Swedish	swe	Indo-European	6.34	0.79	1.82	11.64	5.08
Swahili	swa	Niger-Congo	10.1	0.69	1.54	14.72	5.15
Tamil	tam	Indo-European	6.34	1.04	1.61	12.5	8.12
Telugu	tel	Indo-European	5.87	0.75	1.11	11.64	7.03
Tajik	tgk	Indo-European	6.52	0.77	1.96	13.43	5.39

Thai	tha	Kra-Dai	6.21	1.14	2.56	11.34	4.83
Turkish	tur	Turkic	6.43	0.94	2.09	11.77	6.48
Ukrainian	ukr	Indo-European	6.7	0.78	1.78	10.82	5.87
Urdu	urd	Indo-European	5.34	0.64	0.66	11.17	3.9
Uzbek	uzb	Turkic	7.6	0.99	2.25	11.8	6.58
Vietnamese	vie	Austroasiatic	6.71	1.01	2.33	10.97	4.07
Xhosa	xho	Niger-Congo	9.78	1.27	2.91	12.96	7.19
Yoruba	yor	Niger-Congo	8.46	1.56	3.26	15.48	3.48
Cantonese Chinese	yue	Sino-Tibetan	5.56	0.93	2.07	12.31	3.96
Zulu	zul	Niger-Congo	11.05	1.31	3.03	17.3	7.23

Table 12: Detailed statistics of FLEURS-IPA. (Avg.Phones: average number of phonemes in each word; Avg.Dur.: average duration of each clip).