Cross-Lingual IPA Contrastive Learning for Zero-Shot NER

Anonymous ACL submission

Abstract

Existing approaches to zero-shot Named Entity Recognition (NER) for low-resource languages have primarily relied on machine translation, whereas more recent methods have shifted focus to phonemic representation. Building upon this, we investigate how reducing the phonemic representation gap in IPA transcription between languages with similar phonetic characteristics enables models trained on high-resource languages to perform effectively on low-resource languages. In this work, we propose **CON**trastive **L**earning with **IPA** (**CONLIPA**) dataset containing 10 English and high resource languages IPA pairs from 10 frequently used language families. We also propose a cross-lingual **IPA C**ontrastive learning method (**IPAC**) using the CONLIPA dataset. Furthermore, our proposed dataset and methodology demonstrate a substantial average gain when compared to the best performing baseline.

1 Introduction

One of the facts that links the languages of the world together is shared vocabulary. Languages that are phylogenetically related to one another inherit shared words (cognates) and languages that are in contact with one another borrow words (loanwords) from one another. These etymologically related words tend to share similar meanings and similar pronunciations. Various attempts have been made to leverage this similarity. For example, Bharadwaj et al. (2016) used phonetic feature representations of Uyghur and Turkish to leverage shared names in Named Entity Recognition (NER) and Chaudhary et al. (2018) used IPA (International Phonetic Alphabet) representation to improve NER and machine translation in Bengali (pivoting from Hindi). However, these past approaches have proposed models that learned representations for phoneme strings. Loanwords or cognates have similar embedded representations because their IPA representations are similar. We propose, instead, to learn representations—using contrastive learning—that capture the phonological aspects of etymologically-related words across languages.

Various approaches for zero-shot NER in low-resource languages, where data acquisition is challenging, have been proposed over time. Most previous methods (Yang et al., 2022; Liu et al., 2021; Mo et al., 2024) often employed machine translation with grapheme-based inputs. Since machine translation utilizes prior knowledge of low-resource languages, a method using phonemic representation was proposed for a stricter zero-shot setting (Sohn et al., 2024).

We investigate how reducing the phonemic representation gap in IPA transcription between languages with similar phonetic characteristics enables models trained on high-resource languages to perform effectively on low-resource languages as shown in Figure 1. We selected 10 representative languages from 10 widely spoken language families and collected IPA pairs with English that share the same meaning and similar pronunciation. Using this **CON**trastive **L**earning with **IPA** (**CONLIPA**) dataset, we conducted Cross-lingual **IPA C**ontrastive learning method (**IPAC**) on the phonemic representation space. Extensive experiments and cosine similarity score demonstrate that our method effectively brings the representations of similarly pronounced words across different languages closer together.

Our approach differs from (Sohn et al., 2024) in that the model is explicitly trained to represent IPA in a cross-linguistically meaningful way. It is not merely about token overlap; the model learns to represent phonetically transcribed words in a manner that ensures similarity with etymologically related words, such as named entities, in other languages. (Zouhar et al., 2024) also employed similar tech-

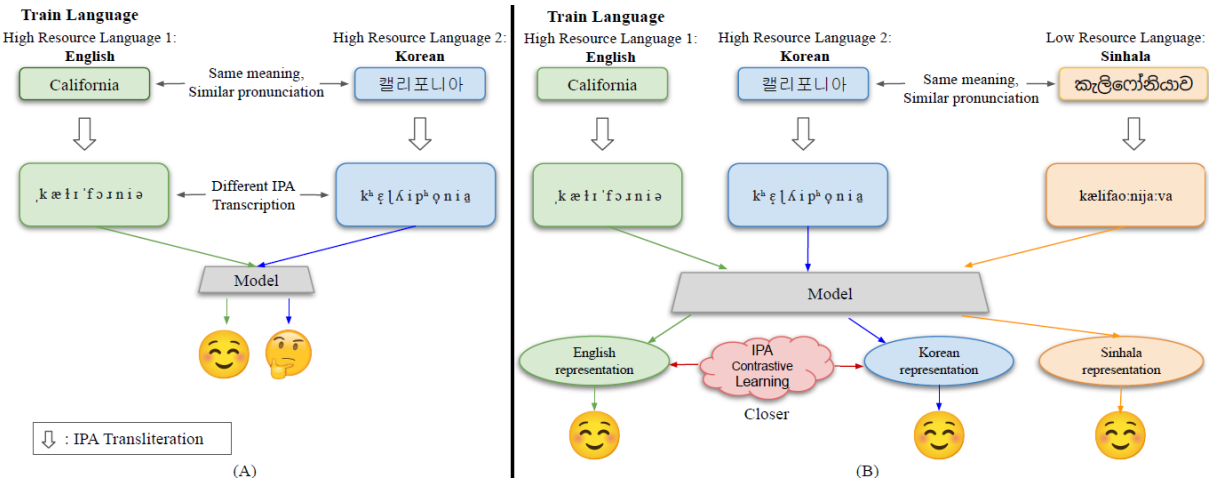


Figure 1: Concept Figure. As shown in (A), existing phonemic models struggle to recognize the same word when IPA representations differ across languages, despite similar pronunciations. In contrast, our method (B) uses IPA contrastive learning to align representations of languages with similar pronunciations, particularly for high-resource languages. This enables effective zero-shot inference for low-resource languages, demonstrating strong generalization.

niques, including metric learning and triplet margin loss, to learn neural representations of IPA strings. However, their approach was monolingual in nature, as both positive and negative samples were drawn from the same language as the anchor, and the metric space was defined based on phonetic features.

We also explored the interesting feature of the Korean language, which allows foreign pronunciations to be recorded using *Hangul* in a way that closely approximates the original pronunciation. Leveraging this feature, we highlight the potential of Korean for future zero-shot NER research.

In general, the main contributions of this paper are as follows:

- We propose the **CON**trastive **L**earning with **IPA** (**CONLIPA**) dataset, which contains IPA pairs of English and 10 languages from 10 widely spoken language families.
- We propose a novel Cross-Lingual **IPA** Contrastive Learning (**IPAC**) approach using the CONLIPA dataset, aimed at reducing the gap in phonemic representations between high-resource languages with similar pronunciations.
- We investigate **Unimodal Contrastive Learning** using exclusively phonemic input, without incorporating multimodal inputs such as images or audio.
- To the best of our knowledge, we are the first to use LLMs, such as ChatGPT, to extract cognate pairs and train a model using these pairs.
- We evaluate the proposed method using WikiANN NER dataset and compare it with baseline methods. Experimental results verify the effectiveness of our method and demonstrate its significant advantages in Zero-Shot NER with low resource language task.

2 Related Work

2.1 Zero-shot Cross-lingual NER

Zero-shot cross-lingual NER is crucial for low-resource languages, where labeled data is scarce. While previous works (Yang et al., 2022; Liu et al., 2021; Mo et al., 2024) used parallel data from machine translation, this approach faces challenges for languages where machine translation is not feasible. ZGUL (Rathore et al., 2023) established a strict zero-shot setting with no target language data, relying on a language adapter trained on typologically similar languages. However, it uses grapheme-based input, limiting its applicability to languages with novel orthographic systems, and is restricted to specific language groups—Germanic, Slavic, African, and Indo-Aryan. In contrast, our approach covers 10 widely spoken language families and does not require overlap between the training and inference languages.

Some works (Bharadwaj et al., 2016; Chaudhary et al., 2018) have utilized phonemic representation for NER, but they did not operate in a zero-shot setting. In contrast, (Sohn et al., 2024) performed

NER by using IPA phonemes as input in a strict zero-shot setting, where no data or prior knowledge was available for the inference language. However, it trained the model exclusively on English data and did not fully address discrepancies in IPA notation for languages with similar pronunciations. The XPhoneBERT(The Nguyen et al., 2023) backbone model used by (Sohn et al., 2024) learns to represent phoneme strings such that similar strings have similar representations. In contrast, our CONLIPA learns to represent phoneme strings, such as names, in a way that ensures they have similar representations to phonologically and semantically related strings in other languages.

2.2 Multimodal and Unimodal Contrastive Learning

Contrastive learning is a self-supervised approach that brings similar data points closer in feature space while pushing dissimilar points apart, enabling the learning of meaningful representations without labeled data by typically using InfoNCE loss (van den Oord et al., 2018; Chen et al., 2020). CLIP (Radford et al., 2021) extends this to learn joint representations of images and text by aligning their features in a shared embedding space, advancing contrastive learning in the multimodal domain, primarily focusing on bridging image-text gaps.

As mentioned in (Huang et al., 2024), unimodal contrastive learning has generally not achieved the same level of success as the unprecedented success of multimodal contrastive learning. The foundational work on contrastive learning has explored key aspects such as alignment and uniformity of contrastive loss (Wang and Isola, 2020), the impact of auxiliary tasks on learning representations (Lee et al., 2021), and optimization perspectives on self-supervised learning (Tian et al., 2020). Additionally, several studies have analyzed contrastive learning in single-modal and multi-view settings (Arora et al., 2019; HaoChen et al., 2021; Tosh et al., 2021; Saunshi et al., 2022). (Wen and Li, 2021) study ReLU networks but differs by requiring an adjustable bias term and not considering multimodal contrastive learning. (Zouhar et al., 2024) also employed related approaches, such as metric learning and triplet margin loss, to learn neural representations of IPA strings. However, their approach was purely monolingual, with positive and negative samples drawn from the same language as the anchor and the metric space defined by phonetic features. Unlike these studies, our approach differs in that it employs a unimodal contrastive learning methodology using only phonemic input based on phonetic features across different languages in a multilingual setting.

2.3 Contrastive Learning with Phoneme Embedding

There have been some research on contrastive learning utilizing phoneme embedding. IPA-CLIP (Matsuhira et al., 2023) is a multimodal method that uses image, text, and IPA, with only using English in both text and IPA. As zero-shot inference experiments on various languages were not conducted, it is difficult to guarantee strong performance across all languages, as IPA symbols may differ between English and other languages.

PLCL (Kewei et al., 2024) is also a multimodal approach that performs contrastive learning between English audio-audio and audio-text pairs. We note that our Cross-Lingual IPA Contrastive Learning (IPAC) clearly differentiates itself by focusing on contrastive learning between phoneme embedding of different languages, rather than within the multimodal domain.

3 CONLIPA Dataset

In this section we provide an overview of how we created the **CON**trastive **L**earning with **IPA** (**CONLIPA**) dataset. The dataset is used in the cross-lingual IPA contrastive learning experiments presented in Section 4.

Language Family	Language	Data
Atlantic-Congo	Swahili	27
Austronesian	Indonesian	86
Indo-European	Hindi	128
Sino-Tibetan	Mandarin	6
Afro-Asiatic	Arabic	34
Austroasiatic	Vietnamese	10
Tai-Kadai	Thai	31
Dravidian	Tamil	71
Turkic	Turkish	52
Koreanic	Korean	7521

Table 1: Selected 10 language families, one of their representative Languages, and the number of data samples per each language.

Language	Target Language Grapheme	English Grapheme	Target Language IPA	English IPA
Swahili	kompyuta	computer	k o m p j u t a	k ə m ˈp j u tɜ˞
Indonesian	Cokelat	chocolate	ʧ o k ˈɛ l a t	ˈtʃɑkə ˌɬeɪt
Hindi	कैमरा	camera	k ɛ̃ː m ɾ ä:	k ɑ ˈm ɛ ɹ ə
Mandarin	沙拉	salad	ʂ a ɭ l a ɬ	ˈs æ ɬ ə d
Arabic	تلفزيون	television	t i l f i z j u: n _ a:	ˈtɛɬə ˌvɪʒən
Vietnamese	vắc xin	vaccine	v ă k ɪ _ s i n ɬ	ˌv æ k ˈs i n
Thai	โฮเทล	hotel	h o: ɬ . tʰ e: l ɬ	h o ʊ ˈtɛ ɬ
Tamil	ஐஸ் கிரீம்	ice cream	ʔ ʌ ɻ s _ k ɨ ɻ i: m	ˈa ɪ s _ ˈk ɹ i m
Turkish	Müzik	music	m y z ˈɪ k	ˈm j u z ɪ k
Korean	어드벤처	adventure	ʌ d ɯ b e n ʧʰ ʌ	æ d v ɛ n ʧɻ

Figure 2: Samples in our CONLIPA dataset for each language.

3.1 Language Selection

We selected 10 major language families and chose one representative language from each family. These languages are high-resource, which makes it easier to obtain IPA pairs with similar phonetic characteristics between the target language and English. We selected the top 9 most widely used language families in the world (*Atlantic-Congo*, *Austronesian*, *Indo-European*, *Sino-Tibetan*, *Afro-Asiatic*, *Austroasiatic*, *Tai-Kadai*, *Drividian*, *Turkic*), and added Korean from the *Koreanic* language family. We included Korean because it is a well-resourced language with a strongly phonemic orthography that, like IPA, has the potential to represent other languages phonemically. This characteristic enabled us to collect a significantly larger amount of data compared to other 9 languages.

Additionally, our CONLIPA dataset used for training contains a minimum of 6 and up to 512 instances per language, enabling efficient and fast fine-tuning. Due to the relatively low computational and memory requirements, the training process incurs minimal computational cost and power consumption, making it more environmentally sustainable. The 10 selected language families, along with the representative languages from each family and the corresponding number of data samples, are presented in Table 1.

3.2 Dataset Creation

We collected pairs of foreign loanwords from English and 10 representative languages that have similar meanings and pronunciations using ChatGPT[1]. Since these languages borrow and use English words directly, the words are transcribed in the closest possible form to original English pronunciation. These words are all loanwords, so it seems that ChatGPT recognizes them as part of a translation task. Additionally, these 10 languages are high-resource languages, meaning that ChatGPT has likely been trained on a large amount of translation data for them. The words obtained were then manually verified by the authors to ensure their pronunciation similarity for each representative language, using Google Translate[2] and online dictionaries. The choice of English as a reference language was motivated by its status as a high-resource language with extensive datasets in NLP, making it likely that models already possess strong representations for English.

As shown in Table 1, the number of such pairs varied significantly across languages. For instance, Mandarin had only 6 pairs due to the limited number of similar pronunciations with English, while Korean, with its ability to represent foreign words phonetically using *Hangul*, allowed for a much larger collection of 7,521 samples. Through experimental evaluation, we found that using **only 512 samples** out of the 7,521 Korean samples yielded the best performance, as shown in Section 6.3.1 Table 4.

We converted the grapheme notation G

[1]https://chatgpt.com/

[2]https://translate.google.co.kr/

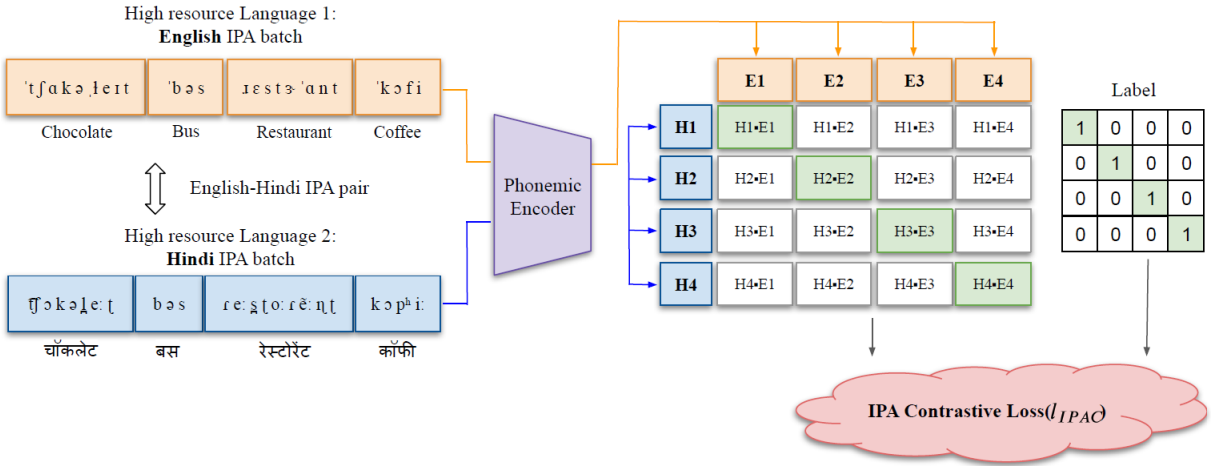


Figure 3: Overall architecture of our IPA Contrastive Learning (IPAC). First, the IPA representations of word pairs with similar pronunciations are obtained from the phonemic encoder for two high-resource languages, such as English and Hindi. Then, these pairs are considered positive pairs, while the remaining samples in the batch are treated as negative pairs to compute the contrastive loss.

of English e and the 10 target languages $t \in \{\texttt{swa}, \texttt{ind}, \texttt{hin}, \texttt{cmn}, \texttt{ara}, \texttt{vie}, \texttt{tha}, \texttt{tam}, \texttt{tur}, \texttt{kor}\}$, into IPA notation I. We used CharsiuG2P toolkit (Zhu et al., 2022) which XPhoneBERT(The Nguyen et al., 2023) originally employed for IPA transliteration. As shown in Figure 2, the dataset format consist of 4 components, which are (G_t, G_e, I_t, I_e).

4 Cross-Lingual IPA Contrastive Learning (IPAC)

Contrastive learning is a widely used self-supervised learning approach, particularly in image-text representation tasks. Its core concept involves training a model to determine whether two input samples are similar or different by evaluating them within a learned latent space.

Our approach differs in that, instead of using image-text pairs of two different modalities, we input IPA transcriptions of two different languages into a phonemic encoder. The goal is to cross-lingually align their phonemic representations. As shown in Figure 3, we performed IPAC by treating pairs of similar-sounding English IPA and target language IPA as positive samples from the CONLIPA dataset, while considering other samples within the batch as negative samples.

We utilized the InfoNCE loss (van den Oord et al., 2018; Chen et al., 2020) in our IPA contrastive learning framework, as it is a widely adopted loss function in contrastive learning. This loss function enhances the mutual information between positive pairs while reducing it between positive and negative pairs. The loss is defined as follows:

$$l(I_e, I_t) = \frac{1}{N} \sum_i -log \frac{\exp\left((I_e^i)^T I_t^i / \tau\right)}{\sum_k \exp\left((I_e^i)^T I_t^i / \tau\right)} \quad (1)$$

where τ is a hyperparameter called temperature coefficient, N refers to the batch size, T refers to the transpose of a matrix and i refers to the i^{th} sample of the batch.

Following the convention, we also perform $l(I_t, I_e)$ symmetrically and calculated the average as shown below, which is used as the final **IPA** Contrastive loss l_{IPAC}.

$$l_{IPAC} = \frac{1}{2}(l(I_e, I_t) + l(I_t, I_e)) \quad (2)$$

5 Experiment Setting

5.1 Models

We followed (Sohn et al., 2024) experimental setting for the three baseline models, mBERT (Devlin et al., 2019), CANINE (Clark et al., 2022) and XPhonebert (The Nguyen et al., 2023). We also compared (Sohn et al., 2024)'s result with ours.

We conducted experiments with models of BERT-base scale: mBERT with 177M parameters, CANINE-C with 132M, and XPhoneBERT with 87,559,687 parameters. Our model initially shares the same number of parameters as the base XPhoneBERT model, as used during pre-training with the WikiANN NER dataset following (Sohn

Case	Input	Model	Languages												AVG	STD
			sin	som	mri	quy	uig	aii	kin	ilo						
CASE 1	grapheme	mBERT	10.71	**44.76**	38.48	55.07	18.70	12.58	**62.37**	79.51					40.27	25.00
	grapheme	CANINE	26.31	43.35	**51.30**	**59.48**	27.19	22.38	54.74	**80.70**					45.68	19.99
	phoneme	XPhoneBERT(baseline)	43.61	38.91	38.07	51.90	44.82	31.03	49.67	73.05					46.38	12.67
	phoneme	CONLIPA(ours)	**45.69**	38.7	39.67	57.7	**45.17**	**34.92**	50.58	73.35					**48.22**	**12.44**
			epo	khm	tuk	amh	mlt	ori	san	ina	grn	bel	kur	snd		
CASE 2	grapheme	mBERT	71.31	16.12	**64.52**	11.90	**63.83**	9.96	48.73	**73.89**	50.44	**83.12**	54.16	35.02	48.58	25.13
	grapheme	CANINE	68.19	27.33	58.07	22.65	61.58	33.53	26.79	68.78	**55.37**	80.07	**57.33**	29.87	49.13	19.86
	phoneme	XPhoneBERT(baseline)	**75.26**	31.86	61.17	44.85	52.58	**40.73**	**59.42**	68.68	49.95	77.61	52.95	47.28	55.20	13.83
	phoneme	CONLIPA(ours)	74.11	**39.95**	60.97	**50.14**	54.03	40.1	53.49	70.73	53.17	72.72	52	**48.44**	**55.82**	**11.62**
			tgk	yor	mar	jav	urd	msa	ceb	hrv	mal	tel	uzb	pan	kir	
CASE 3	grapheme	mBERT	**74.10**	**56.60**	**74.30**	**73.59**	**57.09**	74.98	64.44	**84.93**	**69.94**	**67.24**	**80.04**	**53.98**	**68.14**	**69.18**	**9.28**
	grapheme	CANINE	62.12	51.15	44.28	61.11	42.41	**76.82**	**70.36**	77.51	48.29	37.29	72.54	45.74	57.73	57.49	13.77
	phoneme	XPhoneBERT(baseline)	48.93	50.87	35.12	45.98	33.37	61.76	58.72	58.76	32.52	28.93	60.92	43.85	35.95	45.82	11.85
	phoneme	CONLIPA(ours)	48.19	50.05	38.97	46.24	31.35	62.83	58.16	59.17	39.5	32.57	60.38	49.48	39.37	47.40	10.61

Table 2: Zero-shot F1 score (%) result in **Case 1**, **2**, and **3**. The skyblue boxes indicate better performance compared to the baseline, and the **bold** text represents the best performance for each case and language.

et al., 2024). However, during fine-tuning on the CONLIPA dataset, we incorporated a LoRA adapter and a projection layer. The LoRA adapter adds 1,327,104 parameters, while the projection layer contributes 49,216 parameters, resulting in a total of 88,936,007 parameters. It is important to note that during zero-shot inference, the projection layer is removed, leaving only the LoRA adapter. Given the relatively small size of the LoRA adapter compared to the original XPhoneBERT parameters, this modification resulted in a substantial performance improvement with only a modest increase in model size.

5.2 Dataset

For training, we followed the procedure outlined in Sohn et al. (2024) to train XPhoneBERT on the English WikiANN NER dataset(Pan et al., 2017), which includes seven named entity tags: B-PER, I-PER, B-ORG, I-ORG, B-LOC, I-LOC, and O.

We then fine-tuned the model using our CONLIPA dataset with our IPA contrastive learning methodology. For zero-shot inference, we adopted the settings from (Sohn et al., 2024) for cases 1, 2, and 3. Case 1 includes languages that were not part of the pre-training corpora of mBERT, CANINE, or XPhoneBERT. Case 2 includes only the languages that XPhoneBERT was pre-trained on, while Case 3 includes only the languages that mBERT and CANINE were pre-trained on.

5.3 Implementation Details

For pre-training, we followed the previous approach by setting the max sequence length, as well as both the train and validation batch sizes, to 128. We fine-tuned the pre-trained XPhoneBERT (The Nguyen et al., 2023) from Hugging Face (Wolf, 2019) on the English WikiAnn (Pan et al., 2017) dataset. The training used a learning rate of 1e-5, a weight decay of 0.01, and a warmup ratio of 0.0025.

After obtaining the pre-trained checkpoint, we further performed IPA contrastive learning on our CONLIPA dataset. During this process, we froze the parameters of the original model and added a LoRA adapter with r=8, lora_alpha=32, and lora_dropout=0.1. Additionally, we added a linear projection layer with 64 dimensions, and only the LoRA adapter and projection layer were activated for fine-tuning for 2 epochs. All the other hyperparameters were kept the same as in the pre-training phase.

6 Result and Analysis

6.1 Overall Results

Table 2 compares the overall performance between our method and previous zero-shot NER approaches. It can be observed that our method outperforms the previous phonemic approach (Sohn et al., 2024) in all cases (Case 1, 2, and 3). Additionally, in Case 1, the most stringent zero-shot setting, our model outperformed mBERT (Devlin et al., 2019) and CANINE (Clark et al., 2022) on average.

Notably, in Case 1, which represents a strict zero-shot setting not including any languages used in pre-training, our method shows improved performance in most languages compared to (Sohn et al., 2024). The average performance increases by 1.84%, and the standard deviation decreases in the phonemic contrastive learning condition, indicating more stable and cross-lingually robust results.

Language	Model	Sample Index										
		1	2	3	4	5	6	7	8	9	10	avg
eng-ori	(Sohn et al., 2024)	82.41	60.54	97.17	**74.43**	**88.50**	90.70	86.91	**93.74**	90.46	81.86	84.67
	ours	**90.13**	**62.06**	97.17	73.02	88.00	**92.70**	**90.72**	93.20	**90.78**	**86.09**	**86.39**
eng-khm	(Sohn et al., 2024)	80.72	89.65	88.91	**93.94**	**98.24**	**92.32**	57.05	76.12	69.02	80.05	82.60
	ours	**86.64**	**89.80**	**90.83**	93.67	97.77	92.03	**65.39**	**78.38**	**70.54**	**80.54**	**84.56**

Table 3: Cosine similarity scores(%) for 10 samples of eng-ori and eng-khm pairs.

Model	Languages								AVG	STD
	sin	som	mri	quy	uig	aii	kin	ilo		
XPhoneBERT	43.61	**38.91**	38.07	51.90	44.82	31.03	49.67	73.05	46.38	12.67
Korean-16†	44.62	38.89	38.19	53.59	45.13	31.87	49.59	72.39	46.78	12.40
Korean-32†	44.68	38.82	38.02	55.24	45.08	30.70	49.88	73.10	46.94	12.98
Korean-64†	**45.90**	38.10	38.60	55.69	44.44	33.67	48.25	72.53	47.15	12.33
Korean-128†	45.56	38.49	38.94	54.28	44.48	32.60	47.92	72.22	46.81	12.21
Korean-256†	45.88	37.53	38.73	54.47	44.38	33.93	47.70	72.32	46.87	12.16
Korean-512†	45.69	38.70	39.67	**57.70**	**45.17**	34.92	**50.58**	**73.35**	**48.22**	12.44
Korean-1024†	45.50	36.77	40.91	50.40	42.48	**39.74**	48.62	72.05	47.06	**11.08**
Korean-2048†	45.52	37.14	**41.36**	54.63	42.14	37.46	48.93	72.93	47.51	11.82
Korean-4096†	44.04	33.61	40.14	47.02	40.96	39.83	45.62	70.88	45.26	11.15
Korean-7521†	32.52	25.66	28.52	42.20	36.80	32.16	43.61	64.15	38.20	12.19

Table 4: Ablation study on Korean data number in Case 1. † indicates that the model was trained using all 10 languages of CONLIPA, but with a different number of samples of Korean. The skyblue boxes indicate better performance compared to the baseline, and the **bold** text represents the best performance for each case and language.

6.2 Cosine Similarity of Phonemic Representation

The goal of IPA contrastive learning is to align the cross-lingual representations of languages with similar pronunciations but slightly different IPA transcriptions. To evaluate this, we computed the distance between named entity pairs in English and two low-resource languages, Oriya and Khmer, where each pair has the same meaning, similar pronunciation but different IPA transcription. The distance between the embeddings from each language was calculated using the cosine similarity metric. Figures 5 and 6 in the Appendix present the 10 samples for Oriya and Khmer, respectively.

As shown in Table 3, compared to (Sohn et al., 2024), the results after applying our IPA contrastive learning on both eng-ori and eng-khm showed higher cosine similarity scores in most cases, with the average score also being higher for our method. This demonstrates that our method successfully brought phonemic embeddings with similar meanings and pronunciations closer together across different languages. The t-SNE visualization results for these samples are also provided in the section G of Appendix.

6.3 Ablation Study

6.3.1 Ablation on the number of Korean samples

As can be seen in Table 1, the number of Korean data samples is 7,521, which is significantly higher than that of the other languages. To determine the optimal number of samples for achieving the best performance, we conducted an ablation study by varying the amount of Korean data used in training the model with IPA contrastive learning.

We conducted experiments by gradually increasing the number of Korean data samples, doubling them from 16, 32, 64, ..., up to 7,521, while keeping the data samples of the other 10 languages fixed. As shown in Table 4, the best performance was achieved when the number of Korean data samples was 512. This demonstrates that simply increasing the number of data samples used for IPA contrastive learning does not always lead to better results. While IPA contrastive learning helps bring the representations of similar-sounding words across different languages closer together, excessive usage of it may potentially harm the representations of models pre-trained on original NER datasets. The experimental results of Case 2,3 are also available on Table 9 of Appendix.

Model	Languages								AVG	STD
	sin	som	mri	quy	uig	aii	kin	ilo		
XPhoneBERT	43.61	**38.91**	38.07	51.90	44.82	31.03	49.67	73.05	46.38	12.67
Swahili	44.74	38.71	38.12	53.66	44.89	31.65	49.43	73.24	46.81	12.71
Indonesian	44.43	39.05	39.00	55.53	44.84	32.54	49.43	72.46	47.16	12.39
Hindi	44.62	38.53	38.08	53.69	44.97	30.98	49.28	73.25	46.68	12.85
Mandarin	44.37	**39.2**	38.56	53.61	45.00	31.28	49.63	72.66	46.79	12.53
Arabic	44.46	39.11	38.55	55.02	44.90	32.56	49.4	72.71	47.09	12.44
Vietnamese	44.53	39.07	38.03	55.31	44.95	31.94	50.1	72.69	47.08	12.64
Thai	44.61	39.15	37.94	54.53	**45.25**	31.94	49.89	72.42	46.97	12.48
Tamil	44.43	39.07	37.95	54.68	45.00	30.75	50.01	72.81	46.84	12.01
Turkish	44.62	38.89	38.22	54.93	44.98	30.81	50.09	73.24	46.97	12.96
Korean	44.57	38.51	38.75	55.48	44.86	33.56	49.5	72.5	47.22	12.30
Total	**45.69**	38.7	**39.67**	**57.7**	45.17	**34.92**	**50.58**	**73.35**	**48.22**	12.44

Table 5: Ablation study on each language in case 1. The skyblue boxes indicate better performance compared to the baseline, and the **bold** text represents the best performance for each inference language.

6.3.2 Ablation on each Language

We conducted experiments using only the data from each of the 10 languages in CONLIPA to identify which language performs best when training the model with IPA contrastive learning. As shown in Table 5, Korean achieved the best performance, followed by Indonesian, Arabic, and Vietnamese. However, we can still observe that the *Total* result, using all 10 languages, performed the best, indicating that the data from multiple languages are complementary to each other. The experimental results for Case 2 and Case 3 can also be found in Appendix Table 10.

7 Conclusion

This paper proposes a novel cross-lingual **IPA** **C**ontrastive learning(**IPAC**) methodology to make the phonemic representations of languages with similar pronunciations more similar, aimed at zero-shot cross-lingual NER for low-resource languages. For this purpose, we selected 10 commonly used language families and introduce the **CON**trastive **L**earning with **IPA**(**CONLIPA**) dataset, which includes IPA pairs of similar-sounding words between English and these languages.

Through experiments, we demonstrate that our approach outperforms existing subword, character grapheme-based models, and the basic phoneme-based model. Performance improvements across all cases 1, 2, and 3 confirm the our method's effect on the cross-lingual generalization of phonemic representation, which is crucial for zero-shot NER tasks in low-resource languages where data is scarce.

8 Limitations

Our methodology does not consider all language families worldwide, but rather focuses on 10 language families. Additionally, it is difficult to claim that the representative language selected from each of the 10 language families fully represents all the characteristics of every language within that family. However, our approach demonstrates the potential to improve performance for low-resource languages by leveraging data from high-resource languages, which are relatively easier to obtain.

9 Ethics Statement

In this study, we utilize the publicly available WikiANN dataset (Pan et al., 2017) to train various models across different languages, ensuring that no ethical concerns arise. During the creation of the CONLIPA dataset, we encountered no ethical issues related to its curation or annotation. There were no significant ethical concerns, such as violent or offensive content, and the dataset was used in accordance with its intended purpose.

References

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. 2019. A theoretical analysis of contrastive unsupervised representation learning. In *36th International Conference on Machine Learning, ICML 2019*, 36th International Conference on Machine Learning, ICML 2019, pages 9904–9923. International Machine Learning Society (IMLS). Publisher Copyright: © 2019 International Machine Learning Society (IMLS).; 36th International Conference on Machine Learning, ICML 2019 ; Conference date: 09-06-2019 Through 15-06-2019.

Akash Bharadwaj, David Mortensen, Chris Dyer, and

Jaime Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472, Austin, Texas. Association for Computational Linguistics.

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R. Mortensen, and Jaime Carbonell. 2018. Adapting word embeddings to new languages with morphological and phonological subword representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3285–3295, Brussels, Belgium. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Jonathan H Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

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.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. 2021. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011.

Wei Huang, Andi Han, Yongqiang Chen, Yuan Cao, zhiqiang xu, and Taiji Suzuki. 2024. On the comparison between multi-modal and single-modal contrastive learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Li Kewei, Zhou Hengshun, Shen Kai, Dai Yusheng, and Du Jun. 2024. Phoneme-level contrastive learning for user-defined keyword spotting with flexible enrollment. *arXiv preprint arXiv:2412.20805*.

Bum Jun Kim and Sang Woo Kim. 2025. Temperature-free loss function for contrastive learning. *arXiv preprint arXiv:2501.17683*.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. 2021. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846, Online. Association for Computational Linguistics.

Chihaya Matsuhira, Marc A Kastner, Takahiro Komamizu, Takatsugu Hirayama, Keisuke Doman, Yasutomo Kawanishi, and Ichiro Ide. 2023. Ipa-clip: Integrating phonetic priors into vision and language pretraining. *arXiv preprint arXiv:2303.03144*.

Ying Mo, Jian Yang, Jiahao Liu, Qifan Wang, Ruoyu Chen, Jingang Wang, and Zhoujun Li. 2024. Mclner: Cross-lingual named entity recognition via multi-view contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18789–18797.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Vipul Rathore, Rajdeep Dhingra, Parag Singla, and Mausam. 2023. ZGUL: Zero-shot generalization to unseen languages using multi-source ensembling of language adapters. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6969–6987, Singapore. Association for Computational Linguistics.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. 2022. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pages 19250–19286. PMLR.

Jimin Sohn, Haeji Jung, Alex Cheng, Jooeon Kang, Yilin Du, and David R Mortensen. 2024. Zero-shot cross-lingual NER using phonemic representations for low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13595–13602, Miami, Florida, USA. Association for Computational Linguistics.

Linh The Nguyen, Thinh Pham, and Dat Quoc Nguyen. 2023. Xphonebert: A pre-trained multilingual model for phoneme representations for text-to-speech. In *Interspeech 2023*, pages 5506–5510.

Yuandong Tian, Lantao Yu, Xinlei Chen, and Surya Ganguli. 2020. Understanding self-supervised learning with dual deep networks. *arXiv preprint arXiv:2010.00578*.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. 2021. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206. PMLR.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR.

Zixin Wen and Yuanzhi Li. 2021. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pages 11112–11122. PMLR.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jian Yang, Shaohan Huang, Shuming Ma, Yuwei Yin, Li Dong, Dongdong Zhang, Hongcheng Guo, Zhoujun Li, and Furu Wei. 2022. CROP: Zero-shot cross-lingual named entity recognition with multilingual labeled sequence translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 486–496, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chaoning Zhang, Kang Zhang, Chenshuang Zhang, Trung X Pham, Chang D Yoo, and In So Kweon. How does simsiam avoid collapse without negative samples? a unified understanding with self-supervised contrastive learning. In *International Conference on Learning Representations*.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. Byt5 model for massively multilingual grapheme-to-phoneme conversion. In *Interspeech*.

Vilém Zouhar, Kalvin Chang, Chenxuan Cui, Nate B. Carlson, Nathaniel Romney Robinson, Mrinmaya Sachan, and David R. Mortensen. 2024. PWESuite: Phonetic word embeddings and tasks they facilitate. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13344–13355, Torino, Italia. ELRA and ICCL.

A Language Codes

Table 6 presents the ISO 639-3 language codes for all the languages utilized in the experiments.

Language	ISO 639-3
Amharic	amh
Assyrian Neo-Aramaic	aii
Ayacucho quechua	quy
Cebuano	ceb
Croatian	hrv
English	eng
Esperanto	epo
Ilocano	ilo
Javanese	jav
Khmer	khm
Kinyarwanda	kin
Kyrgyz	kir
Malay	msa
Malayalam	mal
Maltese	mlt
Maori	mri
Marathi	mar
Punjabi	pan
Sinhala	sin
Somali	som
Tajik	tgk
Telugu	tel
Turkmen	tuk
Urdu	urd
Uyghur	uig
Uzbek	uzb
Yoruba	yor
Swahili	swa
Indonesian	ind
Hindi	hin
Mandarin	cmn
Arabic	ara
Vietnamese	vie
Thai	tha
Tamil	tam
Turkish	tur
Korean	kor

Table 6: Language codes for all languages used in the experiments.

Dataset	Lang.	Script	Train	Dev	Test	License
WikiANN	eng	Latn	20k	10k	10k	ODC-BY
	sin	Sinh	100	100	100	
	som	Latn	100	100	100	
	mri	Latn	100	100	100	
	quy	Latn	100	100	100	
	uig	Arab	100	100	100	
	aii	Syrc	100	100	100	
	kin	Latn	100	100	100	
	ilo	Latn	100	100	100	
	epo	Latn	15k	10k	10k	
	khm	Khmr	100	100	100	
	tuk	Latn	100	100	100	
	amh	Ethi	100	100	100	
	mlt	Latn	100	100	100	
	ori	Orya	100	100	100	
	san	Deva	100	100	100	
	ina	Latn	100	100	100	
	grn	Latn	100	100	100	
	bel	Cyrl	15k	1k	1k	
	kur	Latn	100	100	100	
	snd	Arab	100	100	100	
	tgk	Cyrl	100	100	100	
	yor	Latn	100	100	100	
	mar	Deva	5k	1k	1k	
	jav	Latn	100	100	100	
	urd	Arab	20k	1k	1k	
	msa	Latn	20k	1k	1k	
	ceb	Latn	100	100	100	
	hrv	Latn	20k	10k	10k	
	mal	Mlym	10k	1k	1k	
	tel	Telu	1k	1k	1k	
	uzb	Cyrl	1k	1k	1k	
	pan	Guru	100	100	100	
	kir	Latn	100	100	100	
	swa	Latn	1k	1k	1k	
	ind	Latn	20k	10k	10k	
	hin	Deva	5k	1k	1k	
	cmn	Han	20k	10k	10k	
	ara	Arab	20k	10k	10k	
	vie	Latn	20k	10k	10k	
	tha	Thai	20k	10k	10k	
	tam	Telu	15k	1k	1k	
	tur	Latn	20k	10k	10k	
	kor	Hangul	20k	10k	10k	

Table 7: Statistics and license types for the dataset. The table lists the script, number of examples in the training, development, and testing sets for languages in the WikiANN dataset. The dataset is strictly used within the bounds of these licenses.

B Benchmark and License

Table 7 provides information on the datasets, including their statistics and licensing details. Additionally, the CharsiuG2P toolkit (Zhu et al., 2022), used for transliteration, is employed under the MIT license.

C Experimental Result on the trained High Resource Language

The main task of our paper is to perform NER in a strict zero-shot setting, where the inference is conducted on a low-resource language that has never been seen before. However, we also compared the validation set results before and after training on the CONLIPA dataset, which consists of 10 high-resource languages used for Cross-lingual IPA contrastive learning.

As shown in Table 8, in most cases, the performance improved compared to the existing baseline. Although there were occasional instances where the performance dropped below the baseline, the maximum performance improvement was 1.14, while the maximum performance degradation was 0.33. Since the largest performance drop is small, it suggests that performing IPA contrastive learning using the CONLIPA dataset may also be effective in improving the performance of high-resource languages. Additionally, it can be observed that using all 10 languages as *total* shows the best performance both on average and for most individual languages in high-resource languages, too. This suggests that the interaction among 10 representative languages from 10 different language families leads to better results.

D Experimental Result with number of Korean data

We present the ablation study examining the number of Korean instances across all three cases in Table 9.

E Experimental Result with single language

We present the quantitative result of all three cases in Table 10. The method using phoneme representation outperforms in Case 1 and Case 2 in terms of average F1 score(%) and demonstrates more stable results with a lower standard deviation.

F Ablation on the Temperature Coefficient

As discussed in paper (Kim and Kim, 2025), InfoNCE loss (van den Oord et al., 2018; Chen et al., 2020) is commonly employed in contrastive learning since it facilitates learning data representations by capturing the similarities between pairs. While InfoNCE loss plays a crucial role (Wang and Liu, 2021; Zhang et al.), it requires the tuning of a temperature parameter. This critical hyperparameter modifies the similarity scores and governs the intensity of penalties applied to difficult negative samples(Wang and Liu, 2021). This temperature coefficient is represented by τ in equation 1 of Section 4.

To identify an optimal temperature, we conducted an ablation study by varying only the temperature coefficient. The study was performed using 512 Korean data samples, along with data from

Train Language	Zero-shot Inference Language											AVG	STD
	eng	kor	swa	ind	hin	cmn	ara	vie	tha	tam	tur		
XPhoneBERT	76.78	54.88	61.40	64.00	64.99	39.16	55.24	58.49	16.74	59.39	67.28	56.21	16.05
kor	76.69	55.09	61.36	64.02	65.26	38.83	55.22	**58.64**	16.69	59.96	67.18	56.27	16.11
swa	76.79	54.81	61.36	64.07	64.81	39.20	55.14	58.46	16.75	59.28	67.18	56.17	**16.03**
ind	76.8	55.03	61.23	63.98	65.08	**39.22**	55.22	58.53	16.77	59.54	67.34	56.25	16.05
hin	76.74	54.68	61.48	64.04	64.86	**39.22**	55.10	58.51	16.72	59.46	67.15	56.18	16.04
cmn	76.85	55.03	61.31	63.95	65.07	39.14	55.26	58.45	16.76	59.51	67.35	56.24	16.06
ara	76.82	55.00	61.34	64.08	65.13	39.19	55.15	58.57	16.74	59.55	67.28	56.26	16.07
vie	76.87	55.11	**61.52**	64.03	65.21	39.06	55.31	58.52	16.75	59.66	67.36	56.31	16.10
tha	76.87	55.08	61.33	63.99	65.25	39.07	**55.32**	58.51	16.77	59.49	67.43	56.28	16.09
tam	76.82	55.01	61.40	63.96	65.01	39.14	55.27	58.43	16.76	59.38	67.36	56.23	16.06
tur	76.77	54.73	61.45	64.02	64.84	39.19	55.08	58.42	16.75	59.44	67.23	56.17	16.04
total	**76.93**	**55.82**	61.15	**64.09**	**66.13**	38.87	55.25	58.49	**16.87**	**60.49**	**67.55**	**56.51**	16.17

Table 8: F1 score(%) for zero-shot inference on each high-resource language after training on each language of CONLIPA.

10 other languages. As shown in Figure 11, a temperature coefficient of 0.1 yielded the best performance in our experiment.

G Visualization of Phonemic Representation

We analyzed the distance between eng-ori and eng-khm word pairs in Section 6.2 of the main paper. Here, we also visualize the distribution of representations in a zero-shot setting, where phoneme input from a low-resource language is presented solely during inference, without prior exposure during training.

We employed t-SNE to compare how the distribution of representations changes before and after IPA contrastive learning. For this study, we selected Oriya and Khmer as low-resource languages. We used IPA inputs corresponding to 10 English-target language pairs for both Oriya and Khmer, focusing on words with similar pronunciations. The selected words consisted of person, organization, and location named entities. Figures 5 and 6 present the 10 samples for Oriya and Khmer, respectively.

As shown in Figure 4, we compared the t-SNE results before and after IPA contrastive learning using our CONLIPA dataset. The results before learning are shown in (a) and (c), while those after learning are shown in (b) and (d). In the figure, dots of the same color represent pairs of English and target language words with the same meaning, with only the English labels displayed.

In (a) and (c), most of the paired points were distant from each other. Since Oriya and Khmer are low-resource languages, even when input is given in IPA notation, there was a noticeable distance between the paired points. However, in (b) and (d), the distance between these points was significantly reduced.

Note that only the IPA representations of both English and the target language, rather than their grapheme notations, are used for visualization in this process. Additionally, it should be noted that the examples of Oriya and Khmer were not used for any training, such as pre-training with WikiANN or IPA contrastive learning, nor for zero-shot inference. These samples were created and used solely for cosine similarity calculation and t-SNE visualization purposes.

We configured the t-SNE with perplexity=2 and n_iter=300 to generate the visualizations. To ensure a fair comparison, we standardized the axis ranges: for eng-ori, the x-axis range was [-100, 100] and the y-axis range was [-150, 150], while for eng-khm, the x-axis range was [-80, 80] and the y-axis range was [-190, 150].

Case	Model	Languages													AVG	STD
		sin	som	mri	quy	uig	aii	kin	ilo							
CASE 1	XPhoneBERT	43.61	**38.91**	38.07	51.90	44.82	31.03	49.67	73.05						46.38	12.67
	Korean-16†	44.62	38.89	38.19	53.59	45.13	31.87	49.59	72.39						46.78	12.40
	Korean-32†	44.68	38.82	38.02	55.24	45.08	30.70	49.88	73.10						46.94	12.98
	Korean-64†	**45.90**	38.10	38.60	55.69	44.44	33.67	48.25	72.53						47.15	12.33
	Korean-128†	45.56	38.49	38.94	54.28	44.48	32.60	47.92	72.22						46.81	12.21
	Korean-256†	45.88	37.53	38.73	54.47	44.38	33.93	47.70	72.32						46.87	12.16
	Korean-512†	45.69	38.70	39.67	**57.70**	**45.17**	34.92	**50.58**	**73.35**						**48.22**	12.44
	Korean-1024†	45.50	36.77	40.91	50.40	42.48	**39.74**	48.62	72.05						47.06	**11.08**
	Korean-2048†	45.52	37.14	**41.36**	54.63	42.14	37.46	48.93	72.93						47.51	11.82
	Korean-4096†	44.04	33.61	40.14	47.02	40.96	39.83	45.62	70.88						45.26	11.15
	Korean-7521†	32.52	25.66	28.52	42.20	36.80	32.16	43.61	64.15						38.20	12.19
		epo	khm	tuk	amh	mlt	ori	san	ina	grn	bel	kur	snd			
CASE 2	XPhoneBERT	**75.26**	31.86	**61.17**	44.85	52.58	40.73	**59.42**	68.68	49.95	**77.61**	52.95	47.28		55.20	13.83
	Korean-16†	73.48	38.79	59.45	**52.41**	**55.46**	39.91	54.14	70.22	**54.98**	72.48	51.89	48.19		55.95	11.45
	Korean-32†	73.35	38.56	59.00	51.96	55.09	40.22	54.59	70.03	54.55	72.29	52.15	47.96		55.81	11.38
	Korean-64†	73.78	39.74	59.28	52.17	53.76	40.23	53.58	70.37	53.61	72.32	51.89	47.88		55.72	11.39
	Korean-128†	73.66	39.85	59.60	52.08	53.62	40.39	53.38	70.50	54.28	72.45	**52.21**	48.94		55.91	11.30
	Korean-256†	73.79	40.01	60.13	52.01	53.56	41.12	53.68	70.80	53.36	72.38	52.14	48.69		**55.97**	11.28
	Korean-512†	74.11	39.95	60.97	50.14	54.03	40.1	53.49	70.73	53.17	72.72	52.00	48.44		55.82	11.62
	Korean-1024†	73.96	**44.88**	59.86	49.16	51.71	41.15	52.58	**71.05**	51.73	72.80	51.76	**50.02**		55.89	11.03
	Korean-2048†	73.98	41.60	60.37	49.64	52.46	**41.42**	53.53	70.86	52.23	72.59	52.13	48.80		55.80	11.27
	Korean-4096†	72.81	42.18	57.96	50.42	50.78	43.14	50.46	68.43	50.13	71.59	52.33	49.18		54.95	**10.49**
	Korean-7521†	68.01	38.03	56.40	46.82	46.69	39.18	48.62	65.08	45.87	69.23	52.78	45.70		51.87	10.65
		tgk	yor	mar	jav	urd	msa	ceb	hrv	mal	tel	uzb	pan	kir		
CASE 3	XPhoneBERT	48.93	**50.87**	35.12	45.98	33.37	61.76	58.72	58.76	32.52	28.93	60.92	43.85	35.95	45.82	11.85
	Korean-16†	49.01	50.19	38.15	46.19	32.63	61.78	**59.21**	58.95	39.52	32.54	60.70	49.36	38.49	**47.44**	10.56
	Korean-32†	48.90	50.60	37.94	45.99	**33.39**	61.79	58.72	58.72	38.93	32.12	**61.08**	47.56	37.73	47.19	10.60
	Korean-64†	49.22	49.55	38.03	46.20	32.09	62.36	58.25	58.72	39.40	32.17	60.40	48.51	38.34	47.17	10.60
	Korean-128†	**49.56**	49.38	38.54	45.54	32.16	61.81	59.13	59.12	**39.83**	32.69	60.33	48.99	38.81	47.38	10.50
	Korean-256†	48.88	49.50	38.41	46.03	31.70	62.16	58.68	58.90	**39.83**	32.50	60.25	49.42	38.39	47.28	10.58
	Korean-512†	48.19	50.05	**38.97**	46.24	31.35	**62.83**	58.16	**59.17**	39.50	32.57	60.38	49.48	**39.37**	47.40	10.61
	Korean-1024†	45.16	47.22	38.90	**46.63**	28.61	62.57	58.36	58.26	39.64	**33.16**	57.91	51.73	36.86	46.54	10.77
	Korean-2048†	45.87	48.20	38.45	45.77	29.32	62.69	56.68	58.08	39.12	32.13	57.60	**51.77**	38.32	46.46	10.58
	Korean-4096†	45.18	45.81	37.48	45.57	28.17	61.26	53.28	56.11	40.01	31.82	55.50	52.10	36.12	45.26	10.14
	Korean-7521†	31.09	42.08	35.00	44.58	23.80	56.53	50.39	51.34	38.94	30.32	46.19	46.79	35.88	40.99	**9.51**

Table 9: Ablation study on korean data number in Case 1, 2 and 3. † indicates that the model was trained using all 10 languages of CONLIPA, but with a different number of samples of Korean. The skyblue boxes indicate better performance compared to the baseline, and the **bold** text represents the best performance for each case and language.

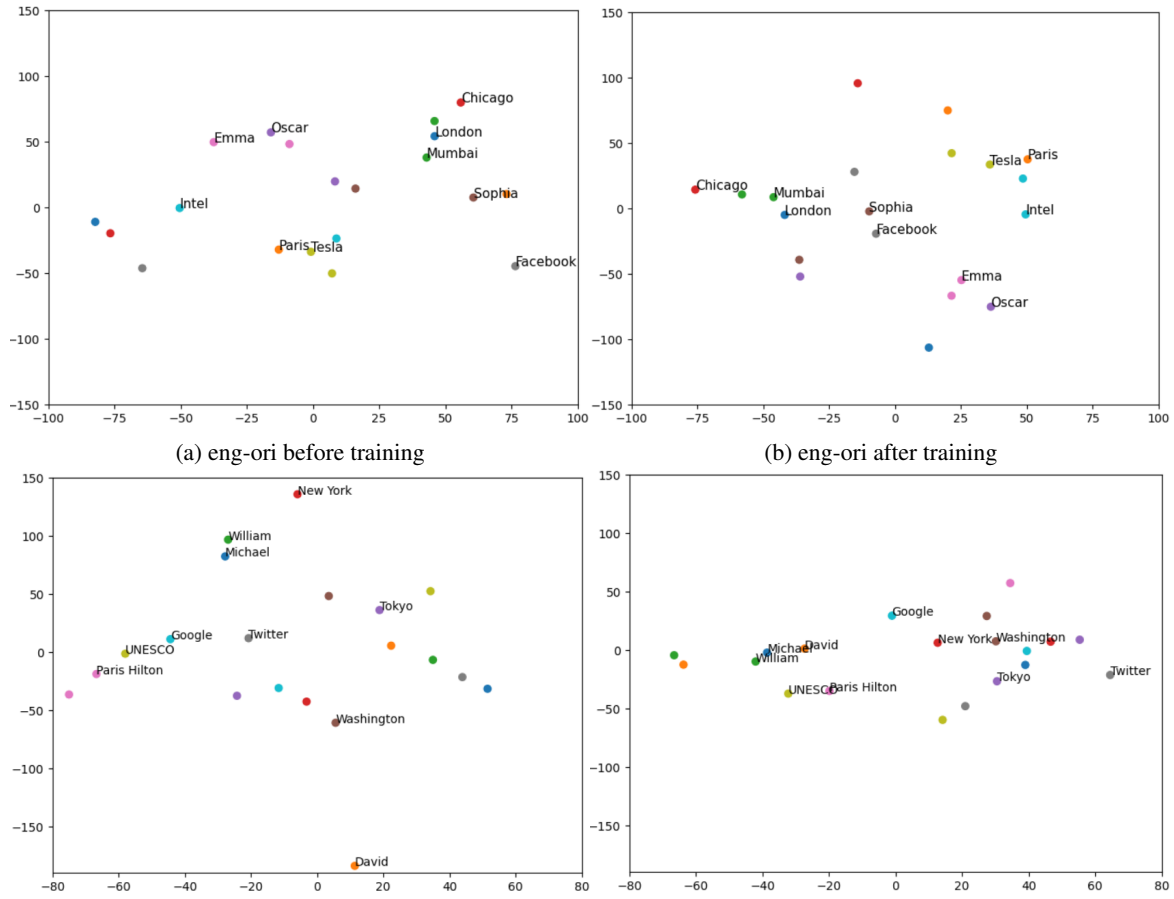
.

Case	Model	Languages													AVG	STD
		sin	som	mri	quy	uig	aii	kin	ilo							
	XPhoneBERT	43.61	38.91	38.07	51.90	44.82	31.03	49.67	73.05						46.38	12.67
CASE 1	Swahili	**44.74**	38.71	38.12	53.66	44.89	31.65	49.43	73.24						46.81	12.71
	Indonesian	44.43	39.05	**39.00**	**55.53**	44.84	32.54	49.43	72.46						47.16	12.39
	Hindi	44.62	38.53	38.08	53.69	44.97	30.98	49.28	**73.25**						46.68	12.85
	Mandarin	44.37	**39.20**	38.56	53.61	45.00	31.28	49.63	72.66						46.79	12.53
	Arabic	44.46	39.11	38.55	55.02	44.90	32.56	49.4	72.71						47.09	12.44
	Vietnamese	44.53	39.07	38.03	55.31	44.95	31.94	**50.10**	72.69						47.08	12.64
	Thai	44.61	39.15	37.94	54.53	**45.25**	31.94	49.89	72.42						46.97	12.48
	Tamil	44.43	39.07	37.95	54.68	45.00	30.75	50.01	72.81						46.84	**12.01**
	Turkish	44.62	38.89	38.22	54.93	44.98	30.81	50.09	73.24						46.97	12.96
	Korean	44.57	38.51	38.75	55.48	44.86	**33.56**	49.5	72.5						**47.22**	12.30
		epo	khm	tuk	amh	mlt	ori	san	ina	grn	bel	kur	snd			
	XPhoneBERT	75.26	31.86	61.17	44.85	52.58	**40.73**	**59.42**	68.68	49.95	77.61	52.95	47.28		55.20	13.83
CASE 2	Swahili	73.34	**38.81**	58.95	51.75	55.16	40.54	54.62	70.00	54.23	72.33	51.76	47.71		55.77	11.35
	Indonesian	73.47	38.71	58.98	**52.25**	55.02	40.04	54.45	70.05	54.88	72.33	51.72	47.87		55.81	11.42
	Hindi	73.3	38.62	58.79	51.61	55.06	40.62	54.79	69.88	54.35	72.32	52.04	47.71		55.76	**11.33**
	Mandarin	73.41	38.62	59.07	51.76	55.31	40.07	54.38	70.07	54.68	72.36	51.63	48.07		55.79	11.43
	Arabic	73.46	38.66	58.99	52.20	55.10	39.97	54.17	70.01	54.85	72.27	51.80	47.76		55.77	11.43
	Vietnamese	73.54	38.56	59.23	51.93	**55.41**	39.94	54.51	70.07	54.94	72.36	52.15	**48.15**		**55.90**	11.45
	Thai	73.48	38.68	59.25	51.84	**55.41**	39.91	54.24	70.1	54.89	72.4	52.13	48.04		55.86	11.45
	Tamil	73.41	38.64	59.19	51.73	55.28	40.12	54.38	70.08	54.88	72.29	52.08	47.94		55.84	11.41
	Turkish	73.30	**38.81**	59.32	51.63	55.13	40.24	55.06	70.01	**55.14**	72.36	52.33	47.34		55.89	11.39
	Korean	73.66	38.73	58.87	51.78	54.8	40.01	54.74	**70.18**	55.04	72.21	51.63	47.15		55.73	11.51
		tgk	yor	mar	jav	urd	msa	ceb	hrv	mal	tel	uzb	pan	kir		
	XPhoneBERT	48.93	50.87	35.12	45.98	33.37	61.76	58.72	58.76	32.52	28.93	60.92	43.85	35.95	45.82	11.85
CASE 3	Swahili	48.35	51.09	37.65	46.04	33.54	61.67	58.66	58.66	38.79	31.78	60.98	47.44	37.36	47.08	10.66
	Indonesian	49.03	50.55	38.13	**46.22**	33.22	62.01	59.04	58.84	39.27	32.22	60.83	**49.13**	37.78	**47.41**	10.62
	Hindi	48.56	51.35	37.65	46.03	**33.56**	61.66	58.35	58.56	38.57	31.78	61.01	47.40	37.42	47.07	10.64
	Mandarin	48.92	50.78	38.11	46.09	33.21	62.00	58.98	58.95	**39.34**	32.21	60.94	48.27	37.94	47.36	10.62
	Arabic	49.12	50.67	38.01	46.17	33.11	61.99	58.94	58.81	39.15	32.20	60.97	48.21	37.96	47.33	10.63
	Vietnamese	**49.20**	50.90	**38.14**	46.01	32.98	62.07	58.81	58.95	39.09	**32.45**	60.92	48.54	**38.11**	47.40	10.62
	Thai	49.05	50.72	38.08	46.02	33.04	62.11	**59.13**	**58.98**	39.17	32.35	61.03	48.47	37.99	47.40	10.67
	Tamil	49.00	50.80	38.05	46.06	33.29	62.00	58.68	58.90	39.14	32.20	61.01	48.19	37.92	47.33	**10.61**
	Turkish	48.26	51.14	37.69	45.97	33.54	61.73	58.37	58.71	38.55	31.85	**61.16**	47.41	37.38	47.06	10.67
	Korean	48.41	**51.48**	37.76	45.57	32.59	**62.25**	57.4	58.71	38.35	31.91	60.54	48.04	37.66	46.97	10.68

Table 10: Ablation study on each language in case 1,2,3.

Case	Temperature	Languages													AVG	STD
		sin	som	mri	quy	uig	aii	kin	ilo							
CASE 1	0.01	37.99	**42.86**	39.96	49.51	**49.15**	27.4	**53.71**	72.49						46.63	13.29
	0.05	46.05	39.14	39.27	54.87	44.96	32.59	49.38	72.02						47.29	12.11
	0.1	45.69	38.7	39.67	**57.70**	45.17	34.92	50.58	**73.35**						**48.22**	12.44
	0.15	**46.62**	36.92	40.10	54.49	44.17	37.74	48.50	72.74						47.66	11.71
	0.2	46.09	36.25	**40.15**	52.45	42.44	38.01	47.84	73.33						47.07	11.88
	0.3	45.73	36.04	40.05	51.32	41.57	38.83	47.28	72.97						46.72	11.70
	0.4	45.62	35.47	38.73	48.45	41.08	39.76	47.8	72.81						46.22	11.68
	0.5	45.71	35.38	38.67	48.8	40.82	39.17	47.8	72.74						46.14	11.75
	0.6	45.82	35.45	39.69	49.35	41.14	39.90	47.56	72.71						46.45	11.57
	0.7	45.71	35.41	38.51	48.68	41.28	39.98	47.81	72.20						46.20	**11.49**
	0.8	45.76	35.27	38.52	48.67	40.41	39.17	48.11	72.07						46.00	11.59
	0.9	45.35	35.34	39.50	49.22	41.07	39.93	47.56	72.63						46.33	11.58
	1.0	46.07	35.24	38.37	48.34	41.16	**39.99**	47.77	72.12						46.13	11.49
		epo	khm	tuk	amh	mlt	ori	san	ina	grn	bel	kur	snd			
CASE 2	0.01	72.95	34.90	59.68	51.80	**57.01**	35.83	**55.04**	**71.36**	53.09	**72.91**	46.56	48.52		54.97	12.92
	0.05	73.70	38.54	59.87	51.11	54.43	39.20	55.01	71.18	**53.75**	72.71	51.58	49.31		55.87	11.76
	0.1	**74.11**	39.95	**60.97**	50.14	54.03	40.10	53.49	70.73	53.17	72.72	52.00	48.44		55.82	11.62
	0.15	74.05	42.78	60.45	51.14	52.32	41.34	52.77	71.22	52.12	72.54	51.95	49.20		55.99	11.14
	0.2	74.08	44.28	60.38	50.84	51.91	42.05	52.57	70.80	51.58	72.61	51.56	49.00		55.97	10.94
	0.3	73.98	45.82	60.26	51.41	51.24	42.98	52.55	70.39	51.14	72.44	51.98	48.78		56.08	10.61
	0.4	73.79	45.82	59.18	52.46	51.59	43.21	52.29	70.77	52.14	72.37	52.98	**49.57**		**56.35**	10.40
	0.5	73.78	45.84	59.15	52.45	51.51	43.21	52.28	70.63	51.82	72.31	53.01	49.26		56.27	10.40
	0.6	73.86	46.17	60.01	51.81	50.93	43.17	51.76	70.30	50.98	72.30	52.02	49.09		56.03	10.52
	0.7	73.85	46.33	60.46	52.40	51.66	42.84	52.34	70.78	50.96	72.36	52.12	48.95		56.25	10.56
	0.8	73.73	45.61	59.16	52.32	51.58	**43.21**	52.24	70.28	51.78	72.32	**53.08**	49.18		56.21	**10.38**
	0.9	73.81	45.91	59.98	51.84	50.90	43.04	51.80	70.27	51.04	72.32	52.15	48.71		55.98	10.56
	1.0	73.83	**46.84**	59.63	**52.46**	51.61	42.91	52.56	70.76	51.31	72.31	52.18	48.93		56.28	10.44
		tgk	yor	mar	jav	urd	msa	ceb	hrv	mal	tel	uzb	pan	kir		
CASE 3	0.01	49.89	**50.81**	38.58	46.08	**33.89**	61.09	58.91	**61.24**	39.41	**33.90**	60.15	48.14	38.13	47.71	**10.35**
	0.05	**50.14**	49.08	38.90	45.63	32.24	62.13	**59.37**	59.74	**40.11**	33.13	**60.58**	49.81	**39.62**	**47.73**	10.50
	0.1	48.19	50.05	**38.97**	**46.24**	31.35	**62.83**	58.16	59.17	39.50	32.57	60.38	49.48	39.37	47.40	10.61
	0.15	47.19	47.85	38.41	46.08	30.27	62.52	58.43	58.57	39.45	32.86	59.42	50.17	38.00	46.86	10.66
	0.2	45.62	47.96	38.03	46.19	30.11	62.44	58.24	57.91	38.95	32.25	59.41	50.98	37.35	46.57	10.79
	0.3	44.99	47.35	38.10	46.23	29.87	62.09	58.13	57.70	38.99	32.35	58.96	50.59	36.90	46.33	10.70
	0.4	45.38	46.90	38.22	45.15	30.32	61.33	57.82	57.88	39.69	32.75	58.89	51.01	37.09	46.34	10.44
	0.5	44.94	47.15	38.13	45.10	30.29	61.31	57.76	57.82	39.62	32.73	58.84	50.93	36.09	46.21	10.52
	0.6	45.24	47.21	38.05	46.23	29.79	61.67	57.92	57.54	39.09	32.37	58.51	51.19	36.67	46.27	10.61
	0.7	45.41	46.80	38.31	46.05	29.76	61.60	58.35	57.79	39.60	32.82	58.36	51.96	36.86	46.44	10.58
	0.8	44.00	46.94	38.15	45.27	30.08	61.21	57.72	57.72	39.80	32.81	58.64	51.39	35.96	46.13	10.52
	0.9	44.95	47.00	38.08	46.10	29.78	61.50	58.00	57.50	39.20	32.34	58.48	51.36	36.09	46.18	10.64
	1.0	45.33	46.83	38.30	46.04	29.68	61.56	58.31	57.75	39.69	32.83	58.25	**52.05**	36.73	46.41	10.58

Table 11: Ablation study on contrastive learning temperature in case 1,2,3.



(a) eng-ori before training

(b) eng-ori after training

(c) eng-khm before training

(d) eng-khm after training

Figure 4: t-SNE (perplexity=2) visualization using 10 eng-ori pairs and 10 eng-khm pairs. Panels (a) and (c) represent the results before IPA contrastive learning, while panels (b) and (d) show the results after learning. Dots of the same color indicate pairs of english and target language words with the same meaning.

English Grapheme	Oriya Grapheme	English IPA	Oriya IPA
London	ଲଣ୍ଡନ୍	ˈlʌndən	lɔɳɖɔn
Paris	ପ୍ୟାରିସ	ˈpɛɹɪs	pjaːrisɔ
Mumbai	ମୁମ୍ବାଇ	məmˈbaɪ	mumbaːi
Chicago	ଚିକାଗୋ	tʃɪˈkɑgoʊ	ʧikaːgo
Oscar	ଓସ୍କାର	ˈɔskɝ	oskaːrɔ
Sophia	ସୋଫିଆ	ˈsoʊfiə	sopʰiaː
Emma	ଏମ୍ମା	ˈɛmə	emmaː
Facebook	ଫେସ୍ବୁକ୍	ˈfeɪsˌbʊk	pʰesbuk
Tesla	ଟେସଲା	ˈtɛsɬə	ʈesɔlaː
Intel	ଇଣ୍ଟେଲ୍	ˈɪnˌtɛɬ	iɳʈel

Figure 5: Ten eng-ori pairs with similar pronunciations

English Grapheme	Khmer Grapheme	English IPA	Khmer IPA
Michael	ម៉ាយកែល	ˈmaɪkəɬ	maaykael
David	ដេវីត	ˈdeɪvɪd	deeviit
William	វិលៀម	ˈwɪɬjəm	viʔliəm
New York	ញូយ៉ក	ˈnu_ˈjɔɹk	ɲuuyɑɑk
Tokyo	តូក្យូ	ˈtoʊkioʊ	toukyou
Washington	វ៉ាស៊ីនតោន	ˈwɑʃɪŋtən	vaasintaon
Paris Hilton	ប៉ារីស ហ៊ីលតន	ˈpɛɹɪs_ˈhɪɬtən	paariih_hiiltɑɑn
Twitter	ទ្វីតធើ	ˈtwɪtɚ	tviitʰəə
UNESCO	យូណេស្កូ	juˈnɛskoʊ	yuuneehkou
Google	គូហ្គល	ˈgugəɬ	kuukɔɔl

Figure 6: Ten eng-khm pairs with similar pronunciations