Towards Language-Agnostic STIPA: Universal Phonetic Transcription to Support Language Documentation at Scale

Jacob Lee Suchardt, Hana El-Shazli and Pierluigi Cassotti
University of Gothenburg
jacob.lee.suchardt@gmail.com
{hana.el-shazli, pierluigi.cassotti}@gu.se

Abstract

This paper explores the use of existing state-of-the-art speech recognition models (ASR) for the task of transcribing speech with narrow phonetic transcriptions using the International Phonetic Alphabet (Speech-to-IPA, STIPA). Unlike conventional ASR systems focused on orthographic output for high-resource languages, STIPA can be used as a language-agnostic interface valuable for documenting under-resourced and unwritten languages. We introduce a new STIPA dataset for South Levantine Arabic and present a large-scale evaluation of STIPA models across 21 language families. Additionally, we provide a use case on Sanna, a severely endangered language. Our findings show that fine-tuned ASR models can produce accurate IPA transcriptions with limited supervision, significantly reducing phonetic error rates even in extremely low-resource settings. The results highlight the potential of STIPA for scalable language documentation and the relevance of training data composition.

1 Introduction and Motivation

Spoken language is a vital channel of communication for the majority of the world's population. In light of the advances in language technology, spoken language can also serve as an intuitive interface, effectively bridging technical hurdles for many non-expert users (Tellex et al., 2020). Consequently, Automatic-Speech-Recognition (ASR) has become a key component for a variety of Natural Language Processing (NLP) applications, such as automatic subtitles, virtual chatbots, and automatic answering machines. In the past decade, ASR – and thus the quality of these applications – has seen vast improvements, with the state-of-the-art (SOTA) ASR systems rivaling and even outperforming humans (Radford et al., 2023; Yadav and Sitaram, 2022).

However, the focus of a majority of ASR systems along with the aforementioned impressive performance (1) only apply to high-resource languages, and (2) rely on Speech-to-Text (STT), i.e. mapping spoken language to standardized orthographic representations.

In this paper, we aim to address these caveats with a practical approach targeted at shifting the focus of ASR from language-dependent orthographical representations to a universal, phonetically motivated notation using the International Phonetic Alphabet (IPA). We achieve this by fine-tuning a SOTA ASR model to predict narrow phonetic transcriptions in IPA, i.e. Speech-to-IPA (STIPA). STIPA models offer a universal pathway from audio to written representation (not necessarily orthography). This audio-to-'text' mapping can serve as an input/output interface for models that operate exclusively on textual data, such as text-based LLMs, which cannot process other modalities. Thus, STIPA models can be particularly beneficial for languages without a standardized orthography and overcome the paradigm centered on high-resource languages. As such, STIPA or phoneme recognition models have also gathered interest as an intermediate step for multi- and cross-lingual orthographic ASR with a subsequent phoneme-to-grapheme conversion stage (Yusuyin et al., 2025; Ma et al., 2025).

Moreover, manual IPA transcription is highly time-consuming, requiring 40–100 hours of human labor per hour of speech—a challenge known as the transcription bottleneck (Seifart et al., 2018), which STIPA models can help to mitigate. Field linguists, particularly those working on under-documented or endangered languages, need phonetic and phonological—rather than orthographic—transcriptions. Orthography often fails to capture pronunciation details crucial for documentation, and many oral languages do not have standardized writing systems.[1]

[1]Even widely spoken languages may not have an established orthography, e.g. Nigerian Pidgin (100M speakers, Lin

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 31423–31439
November 4-9, 2025

Although STIPA has been an area of research for quite some time, more traditional systems have focused on recognizing language-specific phonemes (e.g., Marjou, 2021). Some of these further allow for the subsequent conversion to phonetic symbols through a language-specific phone inventory (e.g., Li et al., 2020; Boulianne, 2022), creating phonemic transcriptions that already encode language-dependent phonological knowledge.

Recently, with the advances of STT-ASR and a growing awareness of the aforementioned bias in NLP to prioritize high-resource languages, research interest in STIPA has been renewed (e.g., Gao et al., 2021; Xu et al., 2021; Taguchi et al., 2023; Zhu et al., 2025). However, the focus is shifting towards cross-lingual or universal phone recognition aimed at predicting language-independent phonetic transcription from speech directly. Given the advantages of multilingual and diverse pretraining demonstrated in ASR (Pratap et al., 2020; Yadav and Sitaram, 2022) and STIPA (Żelasko et al., 2020; Xu et al., 2021; Zhu et al., 2024) the recent SOTA ASR model Whisper (Radford et al., 2023) has remained under-explored for STIPA.

1.1 Contributions:

1. We construct IPA transcriptions for the Arabic Speech Corpus (ASC, Halabi, 2016) by mapping the provided Buckwalter notations to IPA, thus creating a novel STIPA dataset for South Levantine Arabic[2];
2. We demonstrate that Whisper can be fine-tuned for the STIPA task, showing that its latent phonetic representations can be effectively leveraged to go beyond orthographic transcription and produce accurate IPA transcriptions[3];
3. We present a large-scale evaluation with revised evaluation metrics of STIPA models—including the SOTA MultIPA (Taguchi et al., 2023) and our fine-tuned Whisper models—across 21 language families, as well as an evaluation on an extremely low-resource, severely endangered language (Sanna);
4. Our results show that while MultIPA performs well in zero-shot settings, models based on Whisper achieve SOTA results on CommonVoice (CV) seen languages and ASC, and remain competitive on unseen languages—highlighting the advantages of parameter-efficient fine-tuning and leveraging typological similarity.

2 Related Work

This section reviews relevant literature in ASR and STIPA, with a focus on multilingual robustness, zero-shot generalization, and suitability for low-resource scenarios.

ASR In designing a robust ASR pipeline for STIPA, it is essential to evaluate models not only for accuracy, but also for adaptability to multilingual and low-resource scenarios.

Among self-supervised methods, Wav2Vec 2.0 (Baevski et al., 2020) learns contextualized speech features from raw audio using a CNN encoder and Transformer context network, performing well in low-resource settings. Whisper (Radford et al., 2023), a supervised Transformer encoder-decoder trained on 680k hours of multilingual audio (83% English), shows strong general performance, though its multilingual consistency varies (Rouditchenko et al., 2023). It comes in sizes from 39M to 1550M parameters; larger variants offer improved accuracy at the cost of latency (Radford et al., 2023).

Distilled Whisper variants (e.g., distil-large-v2/v3, Kotoba-Whisper) or CrisperWhisper (Zusag et al., 2024) have a mono- or bilingual focus, thus lacking sufficient multilingual support, as is the case for alternative models like Canary (Puvvada et al., 2024), Parakeet (Xu et al., 2023), and State Space Models like Mamba (Shakhadri et al., 2025).

Given Whisper's multilingual strengths and fine-tuning flexibility, it is the most suitable foundation for our STIPA pipeline.

STIPA The STIPA task poses unique challenges in zero-shot and cross-linguistic scenarios, particularly when transcribing under-resourced or undocumented languages.

An early and influential model in this space is Allosaurus (Li et al., 2020), which employs a shared BiLSTM encoder and an allophone projection layer trained with CTC loss to produce phonemic transcriptions. By leveraging Phoible (Moran and McCloy, 2019) phoneme inventories, Allosaurus maps predicted phones to language-specific phonemes across approximately 2,000 languages. Despite this broad phonemic coverage, performance dete-

et al., 2024).

[2]The dataset is available in Zenodo

[3]All the models developed in this work are available in Huggingface and the code is provided in Github.

riorates in zero-shot scenarios of universal phone recognition—for example, the model reports phone error rates (PER) exceeding 80% on languages such as Inuktitut and Tusom. This degradation is likely due to its reliance on training data from only 12 languages.

Building on this, Gao et al. (2021) introduced a Wav2Vec2-based model that integrates typological embeddings from Glottolog and Phoible. This approach enhances phonetic token error rates (PTER) across both seen and unseen languages by enriching the acoustic representations with linguistic context. Similarly, Wav2vec2Phoneme (Xu et al., 2021) fine-tunes Wav2Vec2 on multilingual datasets while employing articulatory-feature-based mappings to support zero-shot phoneme prediction without requiring explicit language labels. This model achieves performance comparable to Wav2vec-U while reducing data requirements by up to 80%, underscoring the utility of large-scale multilingual pretraining in phoneme recognition.

More recently, MultiIPA (Taguchi et al., 2023) has advanced the state of STIPA through an end-to-end approach. Fine-tuned from Wav2Vec2, MultiIPA is trained on seven orthographically transparent languages from CommonVoice 11.0, using high-quality grapheme-to-phoneme (G2P) mappings and a unified IPA vocabulary. The model achieves superior zero-shot transfer performance across typologically diverse, manually transcribed languages, particularly when trained on small but diverse datasets. This work highlights the importance of transcription quality and linguistic diversity over raw data volume in achieving robust STIPA performance. Despite its strengths, MultiIPA exhibits notable limitations, such as the lack of explicit modeling of tonal and suprasegmental phenomena, which are critical in many languages. Additionally, its evaluation is constrained by a limited test set, which may not fully capture the diversity of real-world inputs.

Finally, we highlight the concurrent work of (Zhu et al., 2025) which presents a novel SOTA STIPA model family for broad phonetic transcription, ZIPA, with an efficient Zipformer backbone. Moreover, they introduce IPA-PACK++, an open-source STIPA corpus of more than 17k hours across 88 languages with G2P-based IPA transcriptions, as a refined version of IPA-PACK (Zhu et al., 2024).

3 Data

In this section, we introduce the datasets used for model training and evaluation, and describe the process of creating the dataset for South Levantine Arabic.

CommonVoice (CV) Taguchi et al. (2023) builds on Japanese, Finnish, Greek, Hungarian, Maltese, Polish, and Tamil data from CommonVoice 11.0 (Ardila et al., 2020), a popular, multilingual crowd-sourced corpus consisting of read speech and orthographic transcriptions.

To transliterate the orthographic transcriptions into IPA, Taguchi et al. (2023) used Epitran (Mortensen et al., 2018) for Polish and Tamil (62.78% of training data), and applied their own hand-crafted rules to the other five languages. Taguchi et al. (2023) also refer to the reliable quality of these G2P tools, which they assessed manually, as a prerequisite. The zero-shot test data used to evaluate MultIPA consists of about 100 samples of Hakha Chin, Luganda, Tatar, and Upper Sorbian recordings taken from CommonVoice 11.0. The IPA transcriptions had been created for the study by two human annotators.

VoxAngeles (Chodroff et al., 2024) is a recent corpus based on the UCLA Phonetics Lab Archive (Ladefoged et al., 2009). Building on the CMU UCLA corpus (Li et al., 2021), VoxAngeles adds corrected transcriptions, phone-level alignments, and 1,669 new utterances from 11 languages; for a total of 21 language families and 95 unique languages. While word-level alignments are manually verified, phone-level ones remain unvalidated. VoxAngeles' phonetic diversity, spanning 5,355 samples across 95 languages, and low-resource focus make it valuable for phone recognition and low-resource ASR (Mortensen et al., 2021).

THCHS-30 The THCHS-30 corpus, released by Tsinghua University's CSLT, includes c.a. 35 hours of Mandarin read speech from 40 young speakers recorded in 2000–2001 (Wang and Zhang, 2015). Each speaker read 500 news-based sentences for phonetic coverage (Li et al., 2004). The training set has 10k utterances (25.5 h) from 30 speakers; development and test sets contain 893 (2.3 h) and 2495 (6.3 h) utterances, respectively. Taubert (2023) added phone-level IPA annotations with timing, silence, and punctuation for use as ground truth.

Corpus Creation for South Levantine Arabic (ASC) The Arabic Speech Corpus (ASC), developed by (Halabi, 2016), aims to support speech synthesis and includes ∼4 hours of South Levantine Arabic (Damascian accent) speech across 1813 studio-recorded files. A single native speaker read transcripts sourced from Aljazeera Learn (Al Jazeera Media Network, n.d.) and auto-generated nonsense utterances (896/1780) to optimize phonetic coverage via a greedy algorithm (Halabi and Wald, 2016). Recordings were post-processed for tempo, intensity, pauses, and silences (100 ms); average utterance length is 7.5 s (1–36 s range). Orthographic and phonetic transcriptions in Buckwalter format (Buckwalter, 2002) are included. Buckwalter transcripts were generated from fully diacritized orthographic text using a custom pronunciation dictionary—adapted per source, informed by speaker idiolect, and corrected post-alignment.

We devised a Buckwalter-to-IPA transliteration module to create IPA-based phonetic transcriptions using the tables provided in Halabi (2016, Tables 4-7, pp.46f.) and supplementary information from the Wikipedia entries pertaining to the Buckwalter transliteration[4] and Arabic romanization[5] to resolve missing segments and inconsistencies from the original itself. If diacritized, Arabic is described as having a highly consistent G2P correspondence and Halabi (2016) resolved the few opaque exceptions, e.g., the implicit (unwritten) *Alif* vowel, with a lookup table.

4 Fine-tuning Whisper

As a first step, we verified that the pretrained Whisper tokenizer already includes all common IPA symbols in its vocabulary, eliminating the need for additional tokens to handle unknown characters. Since our approach is language-agnostic—that is, we do not assume prior knowledge of a language's identity nor phone inventory—we use the full prediction space and leave inventory discovery to future work (see Żelasko et al., 2022). To adapt Whisper's decoder, which requires a language ID token, we introduce a new *ipa* token and resize the embedding layer accordingly.[6] All inputs are prefixed with the *ipa* token and the special token signaling the *transcribe* task. We fine-tuned Whisper with and without LoRA adaptation (LoWhIPA/WhIPA, i.e. Whisper-for-IPA), training a total of ten models:

1. WhIPA Base/Large-v2 (trained on CV)
2. LoWhIPA Base/Large-v2 (trained on CV, Mandarin (THCHS-30), South Levantine Arabic (ASC), and their combination (CV+THCHS-30+ASC)

For the Sanna case study, we trained an additional model, referred to as LoWhIPA Large-SR (Sanna Related), using data from related languages: Greek (CV), Maltese (CV), and South Levantine Arabic (ASC). Details on parameter tuning and the decoding strategy are provided in Appendix D and E. In all cases, we subsampled the datasets, using no more than 1,000 examples per language in accordance with Taguchi et al. (2023).

5 Evaluation

This section details the evaluation metrics used, as well as the performance on the CV, THCHS-30, ASC, and VoxAngeles datasets. Among the MultIPA models introduced in Taguchi et al. (2023), this work focuses on MultIPA-1k for comparison, as it is the only publicly released model and the one that achieved the highest cross-lingual transfer performance in the original study. CV WhIPA and LoWhIPA (Base/Large) are directly comparable to MultIPA-1k, having been trained with the same data. Appendix A provides statistics on the number of examples in each dataset, including the training, validation, and test splits.

5.1 Metrics

There are no standardized metrics and benchmarks for evaluating STIPA models. We use both string-level (PER) and phonologically informed (PFER) metrics, with modifications to Taguchi et al. (2023) to improve reliability (see Appendix C).

5.2 Common Voice

The performance across CV seen and unseen languages are reported in Table 1. Whisper-Large models consistently outperform base models, especially for seen languages—LoWhIPA Large achieves the best results (14.18% PER, 4.95% PFER). Larger models generally offer stronger and more stable PER improvements. LoRA fine-tuning improves performance, often surpassing full fine-tuning. For Whisper-Base, LoRA improves PER by 10.59% and PFER by 4.31%. Gains for Whisper-

[4] Buckwalter transliteration

[5] Romanization of Arabic

[6] While reusing an existing language ID token can yield consistent results (Qian et al., 2024), we create a new token to avoid any unintended influence from prior language IDs.

Metric	Model	el	fi	hu	ja	mt	pl	ta	cnh	hsb	lg	tt	Mean
PER	MultIPA	23.96	36.79	36.37	25.21	**18.86**	29.27	37.63	**73.48**	**60.38**	75.99	**62.37**	43.66
	WhIPA Base	50.36	49.90	42.18	40.87	57.38	68.34	61.62	89.30	72.73	84.76	92.26	64.52
	WhIPA Large	9.67	7.75	19.58	11.27	20.55	16.29	21.97	81.61	68.77	73.86	76.32	37.06
	LoWhIPA Base	29.81	46.64	34.84	37.07	45.19	55.48	47.52	82.45	68.51	80.54	75.69	54.89
	LoWhIPA Large	**8.27**	**7.61**	**17.24**	**10.49**	20.12	**15.47**	**20.08**	81.68	65.81	**72.63**	71.90	**35.57**
PFER	MultIPA	7.33	9.27	8.94	7.90	**6.99**	10.20	11.01	**20.08**	**17.56**	**20.57**	**17.25**	12.12
	WhIPA Base	12.74	12.81	13.26	21.90	17.94	26.18	19.12	26.28	21.44	24.02	33.43	20.85
	WhIPA Large	**3.71**	**2.40**	5.35	3.30	7.62	**6.62**	8.03	24.66	22.87	25.61	23.98	11.35
	LoWhIPA Base	7.83	11.20	11.37	19.46	11.80	17.84	14.29	23.04	18.39	20.99	21.81	16.34
	LoWhIPA Large	3.94	2.49	**4.72**	**2.94**	7.54	6.88	**6.57**	25.64	20.11	27.83	21.84	**10.65**

Table 1: Combined results for PER% and PFER% across all models. Mean is computed across all language scores. Vertical line separates seen (left) and unseen (right) test languages.

Large are smaller but still positive. In unseen settings, error rates rise sharply, and model size advantage shrinks—base models sometimes outperform large ones in PFER. Notably, LoWhIPA Base achieves the best mean PFER across the unseen languages (21.06%), nearly matching MultIPA despite minimal training. MultIPA leads in PER for many unseen languages, such as Hakha Chin and Upper Sorbian. Performance on seen data does not predict generalization: LoWhIPA Base excels on unseen PFER despite modest seen-language results, while LoWhIPA Large struggles in cross-lingual transfer. This suggests large models may overfit to training languages, highlighting the limits of seen-language validation. PER and PFER often diverge due to the coarser nature of PER. For example, WhIPA Base has a large PER gap between Greek and Hungarian, but similar PFERs. Likewise, LoWhIPA Base shows comparable PFER for Japanese and Luganda despite a PER gap of nearly 40 points. WhIPA and LoWhIPA perform well on Greek and Finnish in both settings, possibly due to phonological regularity or Whisper's STT pretraining. MultIPA shows consistent PFER across languages, excelling on Maltese. Finally, zero-shot transfer (ZT) models can rival fine-tuned ones (results are reported in Appendix B). ZT-Base outperforms on Hakha Chin and achieves the best PER (63.83%) when excluding outliers. However, ZT PFERs are typically higher, indicating weak overlap between orthographic and IPA transcriptions, and the models are prone to breaking down on unseen languages.

5.3 Monolingual Evaluation on Levantine and Mandarin

The evaluation results on the ASC (South Levantine) and THCHS-30 (Mandarin) corpora are summarized in Table 2. Larger models generally perform better, though exceptions exist in monolingual and cross-lingual settings. Monolingual Levantine models outperform all others on ASC, with Levantine LoWhIPA Large achieving 5.48% PER and 1.56% PFER. Mandarin LoWhIPA Large achieves 2.11% PFER and 33.31% PER on THCHS-30. Tone errors account for only about 1% of PER, while a much larger portion is due to suprasegmental features, mainly duration labels—such as extra-short, half-long, or long. This suggests that tone modeling is strong, but challenges remain in accurately capturing prosody. Cross-lingual CV models achieve performance comparable to MultIPA, particularly excelling in ASC with a PFER of 9.14%. MultIPA offers a balanced trade-off, outperforming CV-based models on THCHS-30 by around 10%, demonstrating robust generalization. The multilin-

Model	THCHS-30		ASC	
	PER	PFER	PER	PFER
MultIPA	**88.52**	**24.27**	49.46	11.48
CV				
WhIPA Base	119.05	49.07	66.09	19.13
WhIPA Large	94.66	34.53	**47.34**	**9.14**
LoWhIPA Base	109.07	41.57	53.87	14.94
LoWhIPA Large	97.16	32.70	50.91	10.90
Levantine				
LoWhIPA Base	129.04	62.86	9.96	2.80
LoWhIPA Large	99.67	78.19	**5.48**	**1.56**
Mandarin				
LoWhIPA Base	43.25	5.67	96.23	30.24
LoWhIPA Large	**33.31**	2.11	95.99	22.05
Combined				
LoWhIPA Base	54.96	9.90	16.32	3.50
LoWhIPA Large	36.40	**2.04**	6.44	1.64

Table 2: Evaluation on THCHS-30 (Mandarin) and Arabic-Speech-Corpus (South Levantine) test sets.

gual model Combined LoWhIPA Large matches or exceeds monolingual models, suggesting that multilingual fine-tuning is effective even with limited data. CV-based models perform worse on THCHS-30, with PERs and PFERs two to three times higher than on ASC. This drop is likely due to phonetic inventory mismatches and Mandarin's linguistic complexity. Several THCHS-30 phones (e.g., [ɔ, ɚ, ɤ]) are absent from CV training inventories. Overall, the low PFERs on ASC across all models affirm the quality and consistency of its phonetic transcriptions. In contrast, THCHS-30 presents more challenges due to detailed prosodic annotations and less training overlap.

5.4 VoxAngeles

Table 3 presents the results on the VoxAngeles dataset. We evaluated models on both seen and unseen languages, with a focus on zero-shot performance. The results for each seen language are presented in Appendix F.

The results reveal that MultIPA achieves the best overall performance among the models, with the lowest PFER at 15.38% and a competitive PER of 60.11%. The CV WhIPA Base model shows a PFER of 32.09% and a PER of 87.80%. The larger Whisper variants demonstrate slightly improved performance, with CV LoWhIPA Large reducing the PER to 66.55% and the PFER to 19.08%. However, the performance gap between the base and large models remains modest, likely due to the absence of Whisper pretraining knowledge on the unseen languages, which diminishes the advantage typically gained from model size.

Related languages form three subgroups: (1) CV-related (Greek, Finnish, Hungarian, Maltese), (2) ASC-related Arabic (aeb, ajp, apc), and (3) Chinese languages (hak, wuu, yue) partially related to THCHS-30. CV-based models perform worse on VoxAngeles than on in-domain CV, though prior fine-tuning still provides benefits. PER degradation is notable; PFER increases are more modest. Hungarian is particularly challenging, with PFERs doubling compared to CV. MultIPA shows robust generalization: it yields the lowest average PER (61.6%) and PFER (15.0%) across the 10 related languages. Combined LoWhIPA Large is the closest baseline. MultIPA also achieves the least degradation on unseen languages.

Levantine models trained on ASC transfer well, especially to Tunisian (aeb) and North Levantine (apc, PFER $< 14\%$). In contrast, Mandarin models trained on THCHS-30 do not transfer well to other Chinese varieties (e.g., Wu PFER $> 40\%$).

Performance on the 85 unrelated languages is most challenging due to diversity, short utterances, and low data overlap. The best base model is CV LoWhIPA Base (PER: 74.65%, PFER: 21.59%). Large multilingual models perform better, with PFERs in the 19–21% range. MultIPA again achieves the best average performance (PER: 59.94%, PFER: 15.43%) across unrelated languages. Model size and multilingual training help transfer, but the composition of training data is key. ASC-trained models also perform well despite being monolingual and recorded under controlled conditions, suggesting quality and consistency matter. Mandarin models degrade significantly on VoxAngeles, often producing overly long outputs, likely since THCHS-30's long-form speech is poorly matched to VoxAngeles' short utterances. In summary, multilingual models like MultIPA are most robust across domains and languages, with typical PFERs of 15–21%. The Chinese group remains an exception and warrants further study.

6 Error Analysis

Assuming STIPA models learn cross-lingual speech representations aligned to a near-universal phone inventory, transcription errors should reflect phonetic proximity or plausible articulatory processes. To test this, we sample single phone mismatch errors from the Combined LoWhIPA Large model predictions. Samples are restricted to reference-hypothesis pairs with identical phone counts and non-zero PER, reducing inclusion of errors unrelated to the actual audio. From 3,534 such samples, we extract the first phone mismatch to minimize alignment-related distortions.

Inspired by Loweimi et al. (2023), mismatched phones are collapsed into broad phonetic classes (BPCs): vowels, stops, fricatives, nasals, affricates, approximants (incl. taps/trills), and a seventh "diacritics" class to distinguish intra-phone differences. Tone mismatches are excluded due to their confinement to THCHS-30.

The results are reported in Figure 1. Most mismatches occur within the same BPC (e.g., approximants to approximants). Diacritic mismatches make up 30–50% of errors per BPC, reflecting phonetic similarity. Vowel mismatches are most common: 1,775 cases (50% of all), with 682 involving diacritics. This pattern replicates Loweimi

Model	CV		Arabic		Chinese		Unrelated		Overall	
	PER	PFER	PER	PFER	PER	PFER	PER	PFER	PER	PFER
MultIPA	**46.8**	**9.3**	**65.1**	14.6	**73.7**	**20.8**	**59.94**	**15.43**	**60.11**	**15.38**
CV										
WhIPA Base	88.7	39.6	84.9	29.5	91.4	33.3	87.70	31.80	87.80	32.09
WhIPA Large	50.3	12.4	75.0	18.6	80.3	25.5	70.98	20.64	70.66	20.45
LoWhIPA Base	61.6	15.3	74.2	20.3	81.2	24.3	74.65	21.59	74.39	21.41
LoWhIPA Large	51.4	14.5	69.7	16.4	78.4	25.0	66.59	19.10	66.55	19.08
Levantine										
LoWhIPA Base	70.2	20.7	65.6	16.3	87.1	28.0	81.99	26.83	81.27	26.34
LoWhIPA Large	67.6	19.6	66.5	18.0	89.3	28.7	79.79	24.87	79.31	24.62
Mandarin										
LoWhIPA Base	169.3	93.7	145.1	68.7	179.7	109.6	175.46	98.81	174.51	98.16
LoWhIPA Large	89.7	21.6	96.3	26.9	108.4	37.1	98.66	30.62	98.63	30.41
Combined										
LoWhIPA Base	69.2	18.0	76.5	18.9	95.2	32.7	81.65	22.94	81.55	23.01
LoWhIPA Large	54.1	11.9	67.5	**12.4**	84.8	25.5	76.76	19.23	75.94	18.99

Table 3: Average PER and PFER scores on VoxAngeles languages related to at least one of the training corpora, with results on novel languages and overall performance in the last columns.

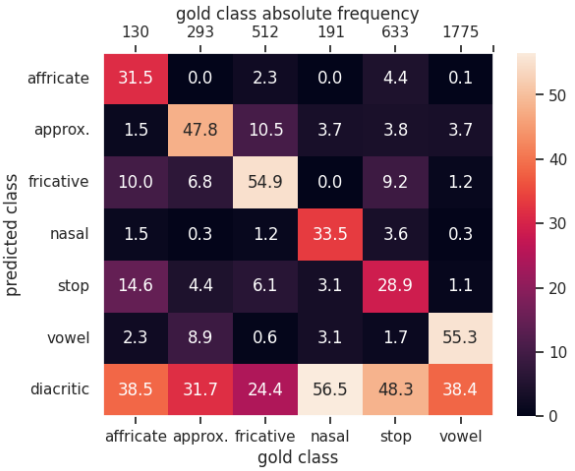

Figure 1: Confusion heatmap of broad phonetic classes (relative frequencies per gold class).

et al. (2023)'s findings and supports the idea that the model attends to perceptual acoustic features. Intra-BPC errors and diacritic mismatches reflect strong phonetic modeling, explaining the common discrepancy between PER and PFER. Following, we provide examples of transcription errors displaying phonetic grounding and not merely text-biases, supporting the model's acoustic alignment.

Example 1 (Greek CV data; 31.58% PER, 9.97% PFER) illustrates that transcription intelligibility remains high despite a 31.58% PER, as many errors are phonetically close or potentially rule-conforming.

(1) *an o kaθenas skeptodan*
an o kafenasketoaðen

Example 2 (Japanese CV data; 5.00% PER, 5.00% PFER) highlights devoicing/assimilation: a dropped /u/ near a voiceless consonant is acoustically plausible.

(2) *joɾoɕikɯoɳegäiɕimäsɯ*
joɾoɕikoɳegäiɕimäsɯ

7 Language documentation: A use case on Sanna

Sanna, or Cypriot Maronite Arabic, is spoken by fewer than two thousand Christian Maronites in Kormakiti, Cyprus. All speakers are bilingual in Sanna and Cypriot Greek, often also knowing Cypriot Turkish, English, or French. This study draws on six audio samples—one from the 1970s (Roth, 1979) (Speaker 3), and five recorded during 2022–2024 fieldwork in Kormakiti (El-Shazli, 2024b,a). Transcriptions were completed by a co-author specializing in phonetic transcription and Sanna. Sociolinguistic data is available for speakers interviewed between 2022–2024. Speaker 1 was aged 76 and the spoken language shows Northern Lebanese Arabic influences, while Speaker 2 was 97-year old and illiterate but orally understood Cypriot Turkish.

Model	S1		S2		S3		Mean	
	PER	**PFER**	**PER**	**PFER**	**PER**	**PFER**	**PER**	**PFER**
MultIPA	67.02	23.11	64.36	20.87	62.27	20.11	64.55	21.36
CV								
WhIPA Base	85.92	30.76	80.53	25.41	81.08	26.10	82.51	27.42
WhIPA Large	77.77	29.68	69.06	22.11	60.50	17.61	69.11	23.13
LoWhIPA Base	79.00	26.78	74.09	21.89	71.05	20.47	74.71	23.05
LoWhIPA Large	70.61	25.65	65.05	**20.18**	56.61	16.54	64.09	20.79
Levantine								
LoWhIPA Base	79.19	27.91	77.45	25.60	71.15	23.07	75.93	25.53
LoWhIPA Large	65.45	23.92	68.39	22.35	61.13	15.20	64.99	20.49
Mandarin								
LoWhIPA Base	118.82	37.41	104.61	32.26	102.09	29.19	108.51	32.96
LoWhIPA Large	97.35	27.75	82.08	24.46	97.28	24.43	92.24	25.54
Combined								
LoWhIPA Base	80.01	27.13	75.61	22.47	64.94	16.87	73.52	22.16
LoWhIPA Large	67.01	23.20	70.38	20.88	56.50	15.60	64.63	19.89
LoWhIPA Large-SR	**63.39**	**21.59**	**63.74**	21.36	**53.56**	**15.03**	**60.23**	**19.33**
Avg.	76.12	24.18	74.47	23.27	69.91	20.16	73.50	22.54

Table 4: STIPA performance for each Sanna speaker. Estimated mean difficulty associated with each speaker is given as the average across all models in the bottom row.

The Sanna results, split by speaker, are shown in Table 4. Alongside average error rates per model, the bottom row estimates each speaker's difficulty via the mean error across all models.

Multilingual and Levantine Large models perform best. Their average PFER hovers around 20%, and PER near 65%, consistent with earlier transfer results. This is notable, as Sanna features spontaneous speech with disfluencies, noise, and a different speaking style. Top performance comes from the LoWhIPA Large-SR model with 60.23% PER and 19.33% PFER. Its design mirrors Sanna's typology, supporting the benefit of typological similarity in cross-lingual transfer. Among Mandarin models, only the large variant (25.54% PFER) approaches LoWhIPA level performance.

MultIPA lags slightly behind LoWhIPA and LoWhIPA Large-SR, especially for Speaker 3, whose 1970s recordings—despite no background noise—have lower audio quality. Yet, non-MultIPA models achieve strong scores here (e.g., 53–61% PER, 15–17% PFER). Speaker 1 sees higher error rates (76.12% PER, 24.18% PFER), possibly due to a persistent background insect noise. Despite clear speech, this noise may challenge STIPA models trained on clean data. Overall, performance is stable across speakers: average scores differ $\leq 6\%$ in both metrics.

Example 3 presents an average transcription sample of Speaker 3 by LoWhIPA Large-SR, as ascertained by scores of 46.15% PER and 16.03% PFER. White spaces were manually added to the model's transcription for readability. The phonetic forms of the reference and hypothesis string bear close resemblance despite a PER of >45%. Moreover, erroneous segments are often phonetically close to the gold transcription (e.g., [x]/[ħ], helping to preserve legibility.

(3) *u istéra bitxaviʃon allik illi flúss ill-ʕarús*
u jistɾɛra biħapiːʃon alːiʔ ili fluis ililʔarut

8 Conclusion

This paper investigates the fine-tuning of Whisper for STIPA using mono- and multilingual data across diverse seen and unseen test settings. We demonstrate that Whisper-Large outperforms Whisper-Base on seen STIPA tasks, achieving state-of-the-art (SOTA) results, though this advantage diminishes in unseen conditions. Fine-tuning with LoRA-based PEFT further improves seen performance, enhances cross-lingual robustness, and reduces computational requirements.

Our findings reinforce previous evidence that multilingual fine-tuning benefits cross-lingual transfer, particularly when fine-tuning and target languages are typologically similar. We also show that high seen STIPA performance can be achieved with as few as 1,000 training samples per language,

supporting previous results from (Taguchi et al., 2023). Additionally, we observe that the number of STT pretraining hours in a language is not a requisite for strong seen or unseen STIPA outcomes.

Our analysis of STIPA fine-tuning in both mono- and multilingual settings under low-resource and low-compute conditions also highlights the efficacy of small, curated datasets combined with PEFT. A custom Buckwalter-to-IPA conversion module is introduced to enrich the ASC dataset, yielding high-quality IPA transcriptions and strong STIPA performance in both seen and unseen evaluations. Furthermore, we explore the ASC and THCHS-30 datasets for fine-tuning, and VoxAngeles for evaluation, extending the scope of STIPA to tonal and suprasegmental phenomena.

Building on these insights, we introduce LoWhIPA Large-SR, a Whisper-Large-v2 model fine-tuned with LoRA on CV Maltese, CV Greek, and ASC South Levantine Arabic data. This model outperforms mono- and multilingual STIPA systems, including the current SOTA MultIPA, in zero-shot transcription of the endangered language Sanna.

Limitations

STIPA research remains in its early stages, as evidenced by the performance gaps across seen, unseen, and cross-lingual settings. Our setup assumed Whisper's encoder learns universal, phonetically rich speech representations. However, this assumption warrants further testing, including the joint fine-tuning of the encoder-decoder or the fine-tuning using different parameter-efficient methods like AdaLoRA. We also left the tokenizer unchanged, letting Whisper implicitly learn IPA tokens. Explicitly restricting the output to IPA tokens could accelerate learning and reduce interference from pretraining. Additionally, neither LoRA settings nor hyperparameters underwent a thorough search. The weak correlation between seen and unseen results suggests both tuning and model selection can be refined. Despite strong seen and cross-lingual performance on ASC and THCHS-30, scaling monolingual data had limited benefits. This suggests that curated, diverse, and clean datasets are more effective. CV fine-tuning data, in particular, could benefit from manual cleaning and consistent IPA transcription standards. Notably, even minimal multilingual input improved over monolingual baselines, supporting few-shot strategies and the inclusion of under-resourced languages—highlighting DoReCo or IPA-PACK++ (Zhu et al., 2025) as valuable resources. STIPA research still lacks a unified evaluation benchmark and universally accepted evaluation metrics which impairs the comparability between previous and concurrent works. Furthermore, although the PFER metric has been revised for this work to capture more phonetic detail, shortcomings – such as an oversensitivity to length discrepancies leading to counter-intuitive scores – remain.

9 Ethical considerations

The recording and phonetic transcription of the Sanna samples were kindly shared with us, and not part of our contributions. We acknowledge the substantial efforts involved and extend our sincere gratitude to both the linguists and the speakers for providing the opportunity to contribute to the documentation of Sanna. We were informed that the original interviews may contain potentially sensitive material. To the best of our knowledge, no such content was included in the excerpts provided for analysis, and every effort has been made to ensure that no such information has been disclosed or published. Apart from the audio recordings and the socio-demographic details referenced in this work, no personally identifying information about the participants was retained.

Our work holds potential for diversifying the field of NLP and supporting the documentation of low-resource languages. However, any work utilizing our assets should be undertaken with the explicit consent of the speakers being transcribed and comply with privacy laws regarding the storage and processing of the contained data. We further urge external research parties to take into consideration the agency and interest of the communities associated with low-resource languages, and to mind spoken language as more than a data resource to commodify.

It must be underlined that the models developed here should not and cannot replace human experts, since their stochastic nature bears the risk of hallucinations and producing otherwise undesirable or unexpected outputs. The recommendation is to use STIPA as a mean of computer-*assisted* transcription, not computer-*based* transcription.

Acknowledgments

This paper builds upon the preliminary work presented by Suchardt (2025). This work has in part been funded by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021). The computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

Al Jazeera Media Network. n.d. Learning aljazeera. https://learning.aljazeera.net/ar.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Gilles Boulianne. 2022. Phoneme transcription of endangered languages: an evaluation of recent ASR architectures in the single speaker scenario. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2301–2308, Dublin, Ireland. Association for Computational Linguistics.

Tim Buckwalter. 2002. Arabic transliteration. http://www.qamus.org/transliteration.htm.

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

Hana El-Shazli. 2024a. Borrowings and suppletion in cypriot maronite arabic. *AIDA Granada: A Pomegranate of Arabic Varieties*, 21:177.

Hana El-Shazli. 2024b. Loanwords in cypriot maronite arabic. *Multilingualism, Variation, Spaces of Literacy. Selected Papers from the 6th International Conference "Crossroads of Language and Cultures" (CLC6)*, pages 445–454.

Sanchit Gandhi. 2022. Fine-tune whisper for multilingual asr with huggingface transformers. https://huggingface.co/blog/fine-tune-whisper.

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

Nawar Halabi. 2016. *Modern standard Arabic phonetics for speech synthesis*. Ph.D. thesis, University of Southampton.

Nawar Halabi and Mike Wald. 2016. Phonetic inventory for an Arabic speech corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 734–738, Portorož, Slovenia. European Language Resources Association (ELRA).

Leopold Hillah, Mateusz Dubiel, and Luis A. Leiva. 2024. "¿te vienes? sure!" joint fine-tuning of language detection and transcription improves automatic recognition of code-switching speech. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, CUI '24. Association for Computing Machinery.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *Preprint*, arXiv:1909.05858.

Peter Ladefoged, Barbara Blankenship, Russell G. Schuh, Patrick Jones, Nicole Gfroerer, Emily Griffiths, Cheryl Hipp, Mayu Kaneko, Gunhye Oh, Keli Vaughan, Sarah Weismuller, Jamie White, WingSze Jamie Lee, Lisa Harrington, Claire Moore-Cantwell, Karen Pfister, Rosary Videc, Samara Weiss, Sarah Conlon, and Rafael Toribio. 2009. The ucla phonetics lab archive. https://archive.phonetics.ucla.edu.

Aijun Li, Zhigang Yin, Tianqing Wang, Qiang Fang, and Fang Hu. 2004. Rasc863-a chinese speech corpus with four regional accents. *ICSLT-o-COCOSDA*.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, Alan W Black, and Florian Metze. 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.

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.

Pin-Jie Lin, Merel Scholman, Muhammed Saeed, and Vera Demberg. 2024. Modeling orthographic variation improves NLP performance for Nigerian Pidgin. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11510–11522, Torino, Italia. ELRA and ICCL.

Wei Ming Liu, Ying Qin, Zhiyuan Peng, and Tan Lee. 2023. Sparsely shared lora on whisper for child speech recognition. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11751–11755.

Yunpeng Liu and Dan Qu. 2024. Parameter-efficient fine-tuning of whisper for low-resource speech recognition. In *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, volume 29, pages 1522–1525.

Erfan Loweimi, Andrea Carmantini, Peter Bell, Steve Renals, and Zoran Cvetkovic. 2023. Phonetic error analysis beyond phone error rate. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3346–3361.

Te Ma, Min Bi, Saierdaer Yusuyin, Hao Huang, and Zhijian Ou. 2025. LLM-based phoneme-to-grapheme for phoneme-based speech recognition. In *Interspeech 2025*, pages 559–563.

Xavier Marjou. 2021. Gipfa: Generating ipa pronunciation from audio. In *Proceedings of the eLex 2021 conference*, pages 588–597.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

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

Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. 2020. Massively multilingual asr: 50 languages, 1 model, 1 billion parameters. In *Interspeech 2020*, pages 4751–4755.

Krishna C. Puvvada, Piotr Żelasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, Jagadeesh Balam, and Boris Ginsburg. 2024. Less is more: Accurate speech recognition & translation without web-scale data. In *Interspeech 2024*, pages 3964–3968.

Mengjie Qian, Siyuan Tang, Rao Ma, Kate M. Knill, and Mark J.F. Gales. 2024. Learn and don't forget: Adding a new language to asr foundation models. In *Interspeech 2024*, pages 2544–2548.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Arlette Roth. 1979. *Le verbe dans le parler arabe de Kormakiti (Chypre): morphologie et éléments de syntaxe*. FeniXX.

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 pretraining for adaptation to unseen languages. *Preprint*, arXiv:2305.12606.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen Levinson. 2018. Language documentation twenty-five years on. *Language*, 94:e324–e345.

Syed Abdul Gaffar Shakhadri, Kruthika KR, and Kartik Basavaraj Angadi. 2025. Samba-asr: State-of-the-art speech recognition leveraging structured state-space models. *Preprint*, arXiv:2501.02832.

Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. 2024. Lora-whisper: Parameter-efficient and extensible multilingual asr. In *Interspeech 2024*, pages 3934–3938.

Jacob Lee Suchardt. 2025. Training for the unexpected: Approaching universal phone recognition for computer-assisted ipa transcription of low-resource languages. Master's thesis, University of Gothenburg. `https://hdl.handle.net/2077/87916`.

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

Stefan Taubert. 2023. Thchs-30 - aligned ipa transcriptions.

Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(Volume 3, 2020):25–55.

Vincenzo Timmel, Claudio Paonessa, Reza Kakooee, Manfred Vogel, and Daniel Perruchoud. 2024. Finetuning whisper on low-resource languages for real-world applications. *Preprint*, arXiv:2412.15726.

Jürgen Trouvain, Jacques Koreman, Attilio Erriquez, and Bettina Braun. 2001. Articulation rate measures

and their relation to phone classification in spontaneous and read german speech. In *Workshop on Adaptation Methods for Speech Recognition*, pages 155–158.

Dong Wang and Xuewei Zhang. 2015. Thchs-30 : A free chinese speech corpus. *Preprint*, arXiv:1512.01882.

Hainan Xu, Fei Jia, Somshubra Majumdar, He Huang, Shinji Watanabe, and Boris Ginsburg. 2023. Efficient sequence transduction by jointly predicting tokens and durations. *Preprint*, arXiv:2304.06795.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2021. Simple and effective zero-shot cross-lingual phoneme recognition. *Preprint*, arXiv:2109.11680.

Hemant Yadav and Sunayana Sitaram. 2022. A survey of multilingual models for automatic speech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5071–5079, Marseille, France. European Language Resources Association.

Yuhang Yang, Yizhou Peng, Hao Huang, Eng Siong Chng, and Xionghu Zhong. 2024. Adapting openai's whisper for speech recognition on code-switch mandarin-english seame and asru2019 datasets. In *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–6.

Saierdaer Yusuyin, Te Ma, Hao Huang, Wenbo Zhao, and Zhijian Ou. 2025. Whistle: Data-efficient multilingual and crosslingual speech recognition via weakly phonetic supervision. *IEEE Transactions on Audio, Speech and Language Processing*, 33:1440–1453.

Jian Zhu, Farhan Samir, Eleanor Chodroff, and David R. Mortensen. 2025. ZIPA: A family of efficient models for multilingual phone recognition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19568–19585, Vienna, Austria. Association for Computational Linguistics.

Jian Zhu, Changbing Yang, Farhan Samir, and Jahurul Islam. 2024. The taste of IPA: Towards open-vocabulary keyword spotting and forced alignment in any language. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 750–772, Mexico City, Mexico. Association for Computational Linguistics.

Mario Zusag, Laurin Wagner, and Bernhad Thallinger. 2024. Crisperwhisper: Accurate timestamps on verbatim speech transcriptions. In *Interspeech 2024*, pages 1265–1269.

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

Piotr Żelasko, Laureano Moro-Velázquez, Mark Hasegawa-Johnson, Odette Scharenborg, and Najim Dehak. 2020. That sounds familiar: An analysis of phonetic representations transfer across languages. In *Interspeech 2020*, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pages 3705–3709. International Speech Communication Association. Publisher Copyright: Copyright © 2020 ISCA; 21st Annual Conference of the International Speech Communication Association, INTERSPEECH 2020 ; Conference date: 25-10-2020 Through 29-10-2020.

A Data statistics

Table 5 presents the statistics of the speech corpora used in fine-tuning and evaluation. It includes the number of audio samples and total duration (in hours or minutes) for the training, development, and test splits, as well as aggregate statistics. The table is organized by dataset source (e.g., MultIPA CV baseline, Arabic Speech Corpus, THCHS-30), and includes additional information such as data filtering criteria and the type of IPA or grapheme-to-phoneme (G2P) conversion used.

B Whisper zero-shot baselines

We evaluate the pretrained Whisper's transcription performance by comparing its outputs to IPA transcriptions on the CV data. This evaluation setting, though uncommon in the literature, is feasible for languages using Latin scripts or reliable Latin transliterations (given the overlap with the IPA). Off-the-shelf Whisper models are used with enforced transcription tasks and language IDs. The first 7 target languages (el–ta) are known to Whisper, though Maltese and Tamil have limited pretraining data (1.26 h and 129.6 h respectively), increasing the chance of transcription errors.

Results in Table 6 exclude non-Latin-script languages (Greek, Japanese, Tamil) from the mean. Two score sets are reported: one including outliers—defined as predictions with >400% PER or >200% PFER—and one excluding them. These outliers, though rare (e.g., only 9 instances or 2.25% among 400 samples in Finnish, Hungarian, Maltese, and Polish for Whisper-Base), significantly skew results, inflating average PER and PFER by c.a. 17.5 points.

Whisper-Large-v2 generally performs better than Whisper-Base, except for Maltese when out-

Language	ID	Training split		Dev. split		Test split		Total		Filter	IPA/G2P
		samples	time	samples	time	samples	time	samples	time	frac.	
MultIPA CV baseline (Ardila et al., 2020)											
Greek	(el)	1 832	1.92h	1 567	1.73h	1 539	1.73h	4 938	5.34h	93%	>6s rules*
Finnish	(fi)	1 739	1.95h	1 408	1.53h	1 284	1.47h	4 431	4.95h	80%	>6s rules*
Hungarian	(hu)	5 659	6.75h	3 634	4.0h	3 554	1.33h	12 847	12.04h	78%	>6s rules*
Japanese	(ja)	5 245	6.4h	3 521	4.0h	3 218	3.83h	11 984	14.19h	77%	>6s rules*
Maltest	(mt)	1 743	2.04h	1 375	1.62h	1 229	1.55h	4 347	5.23h	84%	>6s rules*
Polish	(pl)	11 046	13.04h	6 176	7.23h	6 197	7.37h	23 419	27.6h	71%	>6s Epitran
Tamil	(ta)	10 046	12.49h	3 692	4.42h	4 301	5.14h	18 039	22.05h	28%	>6s, Ca Epitran
MultIPA CV transfer test set (Ardila et al., 2020)											
Hakha Chin	(cnh)							25	1m 30s	-	- human
Upper Sorbian	(hsb)							24	2m 40s	-	- human
Luganda	(lg)							22	1m 57s	-	- human
Tatar	(tt)							29	1m 45s	-	- human
Arabic Speech Corpus (Halabi, 2016)											
South Levantine	-	1 631	3.45h	182	0.37h	100	0.29h	1913	4.10h	-	- Buckw.2IPA
THCHS-30 (Wang and Zhang, 2015)											
Mandarin	-	10 000	25.55h	893	2.3h	2 495	6.31h	13 388	34.16h	-	- Taubert (2023)
VoxAngeles (Chodroff et al., 2024)											
(95 languages)	-							5355	92m 3s	-	- human
Sanna (Roth, 1979; El-Shazli, 2024b,a)											
Speaker 01								52	1m 43s	-	- human
Speaker 02								446	11m 15s	-	- human
Speaker 03								82	2m 46s	-	- human

Table 5: Metadata on fine-tuning and testing corpora. *Custom G2P rules by Taguchi et al. (2023).

Metric	Model	el	fi	hu	ja	mt	pl	ta	cnh	hsb	lg	tt	Mean
PER%	Base	(130.29)	40.74	74.99	(100.25)	81.39	93.32	(139.04)	63.83	158.14	236.77	188.80	116.59
		(107.58)	-	62.13	-	49.19	67.51	(113.93)	-	70.16	64.79	63.84	65.66
	Large	(108.25)	32.74	52.36	(100.00)	87.75	63.48	(121.88)	327.43	79.06	63.29	67.48	96.70
		-	-	-	-	50.22	-	(110.01)	-	83.43	-	-	73.32
PFER%	Base	(69.20)	12.26	37.71	(66.29)	46.56	46.50	(82.13)	23.64	112.7	193.20	148.14	72.12
		(46.69)	-	24.98	-	15.01	21.00	(57.04)	-	25.7	24.12	23.90	24.34
	Large	(47.05)	9.65	22.36	(67.02)	55.97	19.94	(65.00)	278.83	32.8	25.39	24.14	67.39
		-	-	-	-	19.97	-	(53.14)	-	38.51	-	-	30.21
n outliers	Base	1	-	1	-	4	1	2	-	1	3	1	14/800
	Large	-	-	-	-	5	-	1	7	-	-	-	13/800

Table 6: Comparison of zero-shot vanilla Whisper output to IPA transcriptions across train and test languages. The second rows for each model/metric exclude repetition loop outliers. Values in parentheses are excluded from overall means due to non-Latin native scripts.

Metric	Model	el	fi	hu	ja	mt	pl	ta	Mean
PER%	Base	27.09	19.2	33.97	15.09	65.89	47.16	63.78	38.88
	Large	5.26	3.8	7.37	7.95	47.41	9.07	27.92	15.54
PFER%	Base	19.88	6.66	16.23	7.88	39.17	35.52	43.03	24.05
	Large	2.5	1.84	3.15	4.4	27.84	5.32	19.54	9.23

Table 7: Evaluation of zero-shot orthographic vanilla Whisper output with subsequent G2P-based conversion to IPA in comparison to the gold IPA transcriptions generated from the source data's given gold orthographic transcriptions themselves.

liers are removed. Finnish shows strong performance due to its transparent orthography and phoneme-grapheme consistency—Whisper-Large achieves 32.74% PER, indicating about two-thirds correct phone predictions. Given Whisper-Large's reported WER of 14.4% on Finnish CV data, IPA and orthography appear closely aligned here.

For these languages, the G2P-rules used to create the STIPA fine-tuning data can in turn be employed to convert the vanilla Whispers' orthographic outputs to IPA. While the thus achievable error rates (Table 7) are remarkably low in some cases, the rule-based approach has the advantage of comparing standardized transcriptions. This advantage due to standardization also represents a shortcoming however, since dialectal and ideolectal variations that can be present in an utterance are disregarded by the ASR model. Therefore, it is possible that in comparison to G2P-based IPA transcriptions, a STIPA model would score slightly worse than an ASR model with subsequent G2P output transformation, yet be phonetically more accurate. If STIPA results fell well below the orthographcially standardized baselines, it would have been indicative of issues in the training or IPA data quality. Audio-text misalignment is unlikely as evidenced by these zero-shot ASR baselines.

We would like to emphasize that the ASR + G2P approach is only applicable to languages with transparent orthography - that is, languages with a high correspondence between graphemes and phonemes. However, such regularity is not the case for many languages where pronunciation is often irregular (e.g., English, Danish, or French). Orthographic ASR and performant G2P tools are not available for many low-resource and especially endangered languages.

For the unseen set of CV languages (cnh–tt), baselines are more difficult as only Tatar is known to Whisper, and only in limited (14 h) x-to-English training data. To avoid unpredictable behavior from Whisper's automatic language detection, we substitute related Latin-script languages: Hakha Chin → Tibetan, Upper Sorbian → Polish, Luganda → Swahili, Tatar → Turkish. These choices balance linguistic similarity with the goal of producing naïve but interpretable baselines.[7] Since we do not have established G2P rules for these languages, only the pseudo-orthographic outputs can be evaluated with regard to their similarity to the provided phonetic transcriptions.

Whisper-Base outperforms Whisper-Large on three of four unseen languages (Hakha Chin, Upper Sorbian, Tatar), particularly when outliers are excluded. Its PFER is also lower for all but Luganda. Whisper-Large struggles most with Hakha Chin—28% of its predictions are flagged as outliers—while other languages show no such cases.

Compared to seen baselines, unseen results show more frequent breakdowns (1% vs. 6% outlier rate), and higher PER/PFER overall. Still, Whisper-Base yields surprisingly competitive PFERs under outlier exclusion—approaching seen values for Hungarian (23.7%) and Polish (20.5%).

C PER/PFER Metrics

Phone Error Rate (PER) is a common metric for evaluating phone recognition, based on Phone Edit Distance (PED)—a phone-level version of Levenshtein distance. It treats each phone, including those with diacritics, as a single unit, but ignores phonetic similarity, counting all errors equally. Phonetic Feature Error Rate (PFER) extends beyond phone comparisons. Using PanPhon (Mortensen et al., 2016) we map phones to 24-dimensional feature vectors with values in {-1, 0, +1}. We compute PFER using a normalized partial Hamming edit distance, assigning a cost of 1/24 for full feature mismatches, 1/48 for mismatches with undefined features, and 1 for insertions and deletions. The revised implementation of the PER and PFER metrics in this work addresses key shortcomings in the version used by (Taguchi et al., 2023). For PER, improvements include consistent Unicode normalization (e.g., NFD decomposition, diacritic handling), refined tokenization of IPA symbols, and correction for previously mishandled composite phones. The revised PFER metric reintroduces the [+/- syllabic] feature, bringing the phonetic feature vector length back to 24 dimensions as intended, and ensures all feature values remain within the valid {-1, 0, +1} set. Unknown or non-standard phones are handled via a principled fallback system based on PanPhon's diacritic rules rather than naive summation of character vectors, which had previously distorted feature representations. Insertions and deletions are assigned fixed penalties, while substitutions are scored proportionally based on articulatory feature mismatches, yielding a more linguistically informed and stable

[7] Alternative IDs like Burmese or Kazakh were avoided due to frequent non-Latin output.

error estimate.

D Hyperparameters

Across all fine-tuning configurations, we employed fp16 mixed precision, warmup steps equal to 10% of the total fine-tuning steps, and the default AdamW optimizer with linear learning rate decay. The maximum generation length was set to 225 tokens. No hyperparameter search was conducted; instead, hyperparameters were selected based on prior work (Gandhi, 2022; Liu et al., 2023; Hillah et al., 2024; Liu and Qu, 2024; Qian et al., 2024; Timmel et al., 2024). For LoRA fine-tuning, we used r=32, alpha=64, and a dropout rate of 0.05 (Song et al., 2024; Liu and Qu, 2024). A summary of the parameters used is provided in the Table 8. The best checkpoint was selected based on performance on the validation set.

E Decoding strategy

Whisper orthographic STT typically employs 5-beam search, but other strategies have also been applied (e.g., greedy decoding, Yang et al., 2024). STIPA, however, differs from standard STT and targets a more narrow vocabulary. We investigate the impact of beam sizes $n \in \{1, 3, 5, 7\}$ on validation performance using the selected checkpoints of our CV models. Furthermore, we introduce a heuristic to detect repetition-induced outlier predictions which exceed the maximum number of phones that are physiologically possible to articulate within the sample duration, followed by a fallback strategy to reduce their frequency and impact.

We compute each sample's maximum intended speech rate (isr) in (G2P-transliterated) phones/second (ph/s):

$$isr_{\max} = \frac{\sum \text{phones}}{\text{sec}}, \quad \text{sec} = \frac{\text{len(audio array)}}{\text{sampling rate}}$$

Calculated on the filtered validation set of each training language (Figure 2), our $isr_{\max}$ values near alignment with the rates reported in Trouvain et al. (2001) (avg. 13.1 ph/s, max. 21–25 ph/s) and represent realistic speech rate ceilings for read speech.[8]

Our average isr_{max} (dashed line) are slightly lower in comparison, likely because the calculation does not account for pause segments and trailing

[8]The number of phones per sample and the number of BPE tokens required to represent them with the WhisperTokenizer can diverge, depending on phone complexity and token granularity.

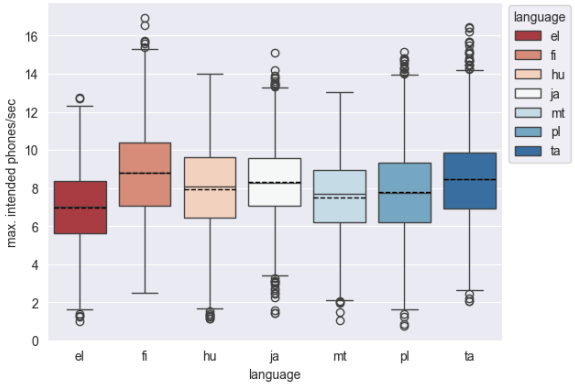

Figure 2: Maximum speech rates per sample as intended phones per second, differentiated by language (validation set).

silences, relying only on IPA transcriptions and total duration. However, since CV samples are short, read-aloud sentences—with participants controlling start/stop—extended silences are rare.

Lenient upper bounds isr_{max_L} are defined by rounding the highest observed isr_{max} per language: el: 13, fi: 17, hu: 15, ja: 16, mt: 14, pl: 16, ta: 17, unseen languages: 20 ph/s. This limit is used to calculate the maximum expected phones per sample:

$$iph_{max_L} = \text{sec} \times isr_{max_L}$$

Predictions exceeding iph_{max_L} trigger decoding fallback strategies and the model retries sequentially with: 1) alternate beam sizes, 2) repetition penalty of 1.15 (Keskar et al., 2019), 3) exponential decay length penalty $\in \{2.0, 3.5, 5.0\}$ for tokens beyond $d = iph_{max_L} \times 0.8$, and 4) forced truncation at iph_{max_L}.

The repetition penalty reduces the likelihood of repeated tokens and is useful against short-loop errors. It may harm longer or repetitive speech, but our maximum input length is known (15 s) and the penalty is only applied to implausible outputs, making adverse effects unlikely.

Fallback option (1) also defines the second core component of our decoding strategy: the primary and backup beam sizes. Validation predictions were run with beam sizes $n \in \{1, 3, 5, 7\}$—both without and with our fallback system. Backup decoding increases the beam size first, then cycles through smaller values.

For simplicity, we assume beam size effects are consistent across checkpoints. Figure 3 shows PFER per checkpoint in all eight conditions (averaged per language; dashed line: macro avg., colored line: medians).

Model	Variant	PEFT	Learn. rate	Batch size	Epochs	Data/Lang.	Lang.	Ckpt
CV								
WhIPA Base	Base	-	1e-5	64	10	1k	7	4
WhIPA Large	Large-v2	-	1e-5	64	10	1k	7	6
LoWhIPA Base	Base	LoRA	1e-3	64	10	1k	7	4
LoWhIPA Large	Large-v2	LoRA	1e-3	64	10	1k	7	8
Levantine								
LoWhIPA Base	Base	LoRA	1e-3	16	10	1k	1	6
LoWhIPA Large	Large-v2	LoRA	1e-3	16	10	1k	1	6
Mandarin								
LoWhIPA Base	Base	LoRA	1e-3	16	10	1k	1	4
LoWhIPA Large	Large-v2	LoRA	1e-3	16	10	1k	1	10
Combined								
LoWhIPA Base	Base	LoRA	1e-3	64	10	1k	9	4
LoWhIPA Large	Large-v2	LoRA	1e-3	64	10	1k	9	6
LoWhIPA Large-SR	Large-v2	LoRA	1e-3	32	10	1k	3	10

Table 8: Fine-tuning hyperparameters and checkpoint choices across baseline, monolingual (Levantine, Mandarin), and Combined (Lo)WhIPA models.

	avg. PFER%/beam size				
Model	**1**	**3**	**5**	**7**	**Chosen**
WhIPA Base	20.27	20.27	20.59	20.38	$5 \rightarrow (7, 3, 1)$
LoWhIPA Base	13.85	13.68	13.43	13.49	$3 \rightarrow (1, 7, 5)$
WhIPA Large	5.84	5.72	5.79	5.69	$3 \rightarrow (7, 5, 1)$
LoWhIPA Large	5.75	5.57	5.52	5.53	$3 \rightarrow (7, 5, 1)$

Table 9: Models' average PFER% on the validation subset using outlier detection and fallback heuristics under variation of beam sizes together with the selection of primary and secondary decoding beam sizes.

Beam size has only minor impact on PFER but the monitoring and fallback heuristics significantly reduce outliers–outliers and means stabilize across WhIPA Base and LoWhIPA Large.

This experiment also informs the choice of primary and fallback beam sizes for each model variant (Table 9, including average PFER per beam size using fallback). Although an initial round of testing—based on an earlier PFER implementation from Taguchi et al. (2023)—guided the strategy selection, minor differences and extensive completed work led us to retain the existing configuration.

F VoxAngeles per language performance

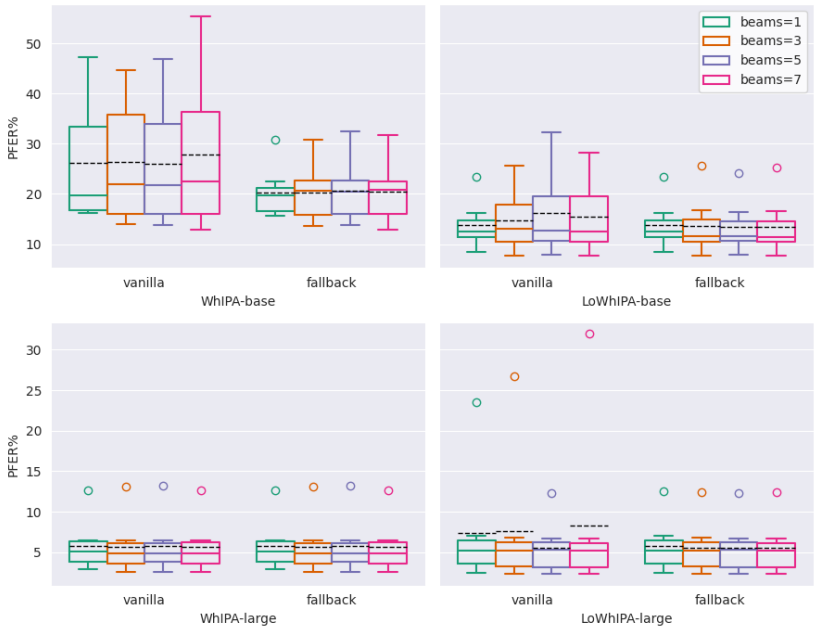

Figure 3: Beam-size dependent performance of model checkpoints on the validation subset and efficacy of speech rate monitoring and fallback heuristics to detect and mitigate repetition loops.

	Model	el	fi	hu	mt	aeb	ajp	apc	hak	wuu	yue	Mean
PER%	MultIPA	**34.7**	45.0	**55.3**	52.1	59.6	**71.8**	64.0	**72.2**	**92.1**	**69.0**	**61.6**
	CV											
	WhIPA Base	65.5	61.8	160.5	67.2	102.6	73.9	78.2	94.5	104.6	77.9	88.7
	WhIPA Large	36.3	45.0	67.6	52.3	69.4	87.9	67.6	76.0	105.1	72.1	67.9
	LoWhIPA Base	51.1	62.0	69.1	64.0	72.6	75.4	74.7	79.1	97.9	75.7	72.2
	LoWhIPA Large	35.9	**40.0**	74.8	54.8	64.8	79.0	65.3	76.4	100.0	70.8	66.2
	Levantine											
	LoWhIPA Base	64.0	63.6	93.0	60.3	61.2	76.5	59.0	88.8	103.3	81.3	75.1
	LoWhIPA Large	64.6	51.7	97.0	57.0	**56.2**	86.3	**56.9**	91.5	105.9	84.7	75.2
	Mandarin											
	LoWhIPA Base	176.7	139.4	211.2	150.1	144.1	108.7	182.6	100.6	255.6	196.1	166.5
	LoWhIPA Large	94.7	73.0	100.5	90.7	96.6	98.1	94.0	103.2	115.9	116.1	98.3
	Combined											
	LoWhIPA Base	54.8	61.8	86.6	73.6	65.1	96.0	68.3	94.0	116.7	89.4	80.2
	LoWhIPA Large	43.9	49.9	70.2	52.5	61.6	78.8	62.0	84.6	111.5	74.1	68.9
	LoWhIPA Large-SR	40.4	45.9	93.4	**48.3**	57.9	82.1	60.0	86.4	106.4	89.3	71.0
PFER%	MultIPA	**4.1**	**8.2**	**16.3**	**8.7**	16.7	**14.0**	13.3	**19.5**	**34.4**	14.4	**15.0**
	CV											
	WhIPA Base	19.9	16.9	100.8	20.7	44.9	22.1	21.4	33.1	44.8	20.9	34.6
	WhIPA Large	4.6	13.9	19.1	12.1	16.0	28.2	11.6	25.7	43.2	**14.3**	18.9
	LoWhIPA Base	11.6	17.3	18.3	13.8	16.7	26.9	17.3	20.8	38.3	18.1	19.9
	LoWhIPA Large	5.7	12.4	29.2	10.8	16.5	21.1	11.6	23.7	43.1	14.4	18.9
	Levantine											
	LoWhIPA Base	12.6	18.8	37.3	14.0	17.3	18.8	12.9	25.6	45.0	19.3	22.2
	LoWhIPA Large	10.5	12.8	40.8	14.4	15.1	27.9	11.0	27.0	43.4	22.0	22.5
	Mandarin											
	LoWhIPA Base	102.7	63.3	131.0	77.8	69.6	30.8	105.6	38.6	184.9	122.2	92.6
	LoWhIPA Large	14.9	14.2	37.6	19.9	29.6	26.1	25.1	32.4	46.5	40.7	28.7
	Combined											
	LoWhIPA Base	9.5	12.7	28.7	21.1	11.2	31.9	13.7	26.9	51.8	28.5	23.6
	LoWhIPA Large	5.8	11.7	19.5	10.4	**9.5**	16.9	10.6	22.7	46.0	16.3	17.0
	LoWhIPA Large-SR	8.3	15.8	38.9	9.9	14.4	23.6	**8.8**	33.7	50.9	24.5	22.9

Table 10: Individual PERs and PFERs on VoxAngeles languages related to at least one of the training corpora, divided by language group.