WikIPA: Integrating WikiPron and Lingua Libre for Multilingual IPA Transcription

Pierluigi Cassotti[1], Jacob Lee Suchardt[2], Domenico De Cristofaro[3]
[1]University of Gothenburg, [2]Leipzig University & ScaDS.AI Dresden/Leipzig, [3]Free University of Bozen
pierluigi.cassotti@gu.se, jacob.lee.suchardt@gmail.com, ddecristofaro@unibz.it

Abstract

We present WikIPA, a new multilingual benchmark designed for automatic speech-to-IPA (STIPA) transcription. By integrating human-curated IPA transcriptions from WikiPron with spoken recordings and metadata from Lingua Libre, WikIPA connects textual phonetic representations with real speech across 78 languages. This open resource supports both broad (phonemic) and narrow (phonetic) transcription tasks, enabling fine-grained evaluation of multilingual phonetic transcription systems. WikIPA provides over 289,000 paired entries and serves as a large-scale foundation for STIPA. We benchmark several state-of-the-art STIPA systems, including MultIPA, (Lo)WhIPA, and ZIPA. Results show that ZIPA achieves the lowest mean error rates across most languages, outperforming Whisper- and Wav2Vec-based baselines. Error analyses reveal that remaining discrepancies largely stem from minor phonetic confusions rather than complete transcription failures, emphasizing the challenge of modeling fine-grained articulatory variation. WikIPA thus establishes the first systematic, multilingual evaluation framework for speech-to-IPA transcription and highlights the potential of combining open, community-driven resources to advance STIPA evaluation.

Keywords: multilingual speech datasets, IPA, speech-to-IPA

1. Introduction and Motivation

The *International Phonetic Alphabet* (IPA) is the internationally recognized standard for representing the sounds of spoken language (International Phonetic Association, 1999). By encoding speech sounds in a transparent manner, the IPA provides a universal descriptive framework for documenting pronunciation across languages and speakers.

IPA plays a pivotal role in both linguistic research and applied speech sciences, ranging from speech disorder assessment to second-language pronunciation learning. In speech-language pathology, it facilitates accurate diagnosis and monitoring of speech sound disorders by capturing fine articulatory details (e.g., [θ] for [s]) (Ball and Rahilly, 2014). In language learning, the IPA helps learners distinguish subtle phonetic contrasts (e.g., [ɪ] vs. [iː]) and supports computer-assisted pronunciation training and automatic pronunciation evaluation (Zhou et al., 2025).

Recent advances in data-driven methods have made it increasingly feasible to develop speech-to-IPA (STIPA) transcription systems that convert raw speech audio directly into sequences of IPA symbols. Unlike traditional Automatic Speech Recognition (ASR) systems that rely on standardized orthography, STIPA models aim to capture the actual articulatory and acoustic realization of speech sounds.

Despite this promise, most existing STIPA resources and benchmarks remain dominated by grapheme-to-phoneme (G2P) approaches. A G2P system predicts phonemic representations from orthographic input, learning statistical mappings between spelling and sounds from a pronunciation lexicon. In many cases, such systems are applied to traditional ASR datasets, such as Common Voice (CV) (Ardila et al., 2020), to automatically generate IPA transcriptions, producing resources that reflect orthography-based phonemic regularities rather than the fine-grained phonetic variation present in real speech. G2P models are inherently constrained by the structure and availability of written language, forcing the STIPA task to approximate *phonemic* forms, i.e., canonical pronunciations abstracted from orthography, rather than the *phonetic* detail realized in speech.

This orthography-based paradigm has several key drawbacks. First, it presupposes the existence of a standardized writing system and large pronunciation dictionaries, excluding thousands of low-resource or unwritten languages. Second, G2P accuracy depends heavily on grapheme-phoneme correspondence, which varies widely across languages: mappings are relatively transparent in Finnish or Turkish, but highly irregular in English, French, or Danish. Finally, G2P transcriptions cannot capture fine-grained articulatory variations (e.g., coarticulation, allophony, or diacritics), which often occur in spontaneous speech.

Since manual phonetic transcription has a high time cost, many STIPA corpora are in fact built from G2P-generated or semi-automatic transcriptions, providing only approximate phoneme-level information. This limits the development and evaluation of true STIPA models that aim to predict detailed IPA sequences from real speech.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8465–8475
11–16 May 2026.

To address this gap, we introduce **WikIPA**[1], a new multilingual benchmark designed to support the STIPA task. WikIPA integrates two large-scale, community-driven resources: **WikiPron** (Lee et al., 2020), which provides human-authored IPA transcriptions extracted from Wiktionary, and **Lingua Libre**[2], a Wikimedia project that collects spoken word recordings. By integrating these complementary datasets, WikIPA links verified phonetic transcriptions with corresponding speech audio across 78 languages, encompassing both broad (phonemic) and narrow (phonetic) IPA forms when available. Despite the IPA transcriptions from WikiPron not being aligned with the specific speech realizations in Lingua Libre, a limitation shared by prior approaches, WikIPA substantially advances the field by providing extensive, human-curated IPA annotations for multilingual STIPA research. We further (1) conduct the first large-scale evaluation of state-of-the-art STIPA models (MultIPA, ZIPA, and (Lo)WhIPA) on WikIPA, finding that ZIPA consistently achieves the lowest error rates, and (2) perform a manual error analysis revealing that most model errors arise from minor phonetic confusions.

2. Related work

2.1. STIPA Datasets

While a number of STIPA datasets are currently freely available, they mainly fall into one of two categories: i) small datasets with high-quality, manually audited transcriptions, which are insufficient to uphold machine-learning, and ii) large scale datasets with G2P-based IPA transcriptions to which the aforementioned drawbacks apply.

The monolingual **Tusom2021** (Mortensen et al., 2021) dataset, whose IPA transcriptions were created and audited manually, but amounts to just under an hour of audio, despite the estimated 200 hours of effort, falls into the first category.

VoxAngeles (Chodroff et al., 2024) presents great linguistic variety and high-quality, manual IPA transcriptions. The recordings comprise single word recordings across 95 languages (21 language families). The release improves on the original (Ladefoged et al., 2009) and subsequent (Li et al., 2021) iterations with audited phonetic transcriptions and phone-level alignment. However, at ∼1.5 hours total recording time, this dataset is better suited for STIPA evaluation purposes.

The **Arabic Speech Corpus** (ASC) (Halabi, 2016) contains about 4 hours of South Levantine Arabic from a controlled, single speaker setting. The dataset was enriched with IPA transcriptions in Suchardt et al. (2025) through a semi-automatically generated Buckwalter transliteration.

Similarly, the **THCHS-30 database** (Wang and Zhang, 2015) contains controlled recordings of Mandarin speech (∼34 hours). The dataset can be used for STIPA with the IPA transcriptions by Taubert (2023), which were created by mapping Pinyin from a pronunciation dictionary to IPA symbols but contain (generalized) phonetic detail, such as lexical tones and duration markers.

Taguchi et al. (2023) put forward resources to create IPA transcriptions for a portion of **Common-Voice** (CV) – Japanese, Finnish, Greek, Hungarian, Maltese, Polish, and Tamil – using a mix of Epitran (Mortensen et al., 2018) (Polish, Tamil) and custom G2P rules. While data for the previous two sources stem from supervised speaker recordings, CV data are crowd-sourced recordings. Therefore, dialectal or idiolectal variety, while beneficial for Speech-to-Text training, could be contained and lower the usability for STIPA.

Most recently, Zhu et al. (2025) introduced **IPAPACK++**, a novel version of IPAPACK Zhu et al. (2024), which normalized character encodings, removed non-IPA fragments, and integrated data from numerous multi- and monolingual ASR sources. At 17,132 hours and 88 languages, IPAPACK++ presents the largest collection of paired speech and IPA transcriptions. Transcriptions were created automatically using rule-based (Epitran) and predictive (CharsiuG2P; Zhu et al., 2022) G2P systems with a focus on broad transcriptions. In accordance with the outlined shortcomings of G2P-based STIPA learning, Zhu et al. (2025) find that even advanced STIPA models tend to predict transcriptions that conform to the languages' "standard" variety and do not account for sociophonetic variance. They conclude that massive scales of G2P-based STIPA data are likely insufficient for universal phone recognition.

2.2. STIPA Models

In recent years, a number of end-to-end STIPA models have been proposed that were attributed cross-lingual capabilities. In particular, insights from Allosaurus (Li et al., 2020), Gao et al. (2021), and Wav2vec2Phoneme (Xu et al., 2021) were instrumental in the development of the more recent models/model families which we evaluate in this paper.

MultIPA[3] (Taguchi et al., 2023) is a fine-tuned XLSR-53 model (Conneau et al., 2021) built on Wav2vec2 (Baevski et al., 2020). It was trained on the CommonVoice 11.0 subset described in Taguchi et al. (2023) and outperformed Allosaurus

[1]The dataset is available on HuggingFace.

[2]https://lingualibre.org/

[3]https://github.com/ctaguchi/multipa

and Wav2vec2Phoneme in the cross-lingual transfer setting on Phonetic Feature Error Rate (PFER).

ZIPA[4] models, built on the Zipformer architecture, were introduced in Zhu et al. (2025) and trained on the accompanying IPAPack++ dataset. Of the presented models, we use the transducer variants ZIPA-T-SMALL/LARGE (65M/302M parameters) and ZIPA-CR-NS-SMALL/LARGE (64M/300M parameters), named after the Consistency-Regularized (CR) Connectionist Temporal Classification (CTC) loss and Noisy Student (NS) methods used during training. ZIPA models, especially ZIPA-T and LARGE variants, exhibit competitive or superior performance to Allosaurus, Wav2vec2Phoneme, MultIPA, and Whisper-PPT – a Whisper-small model (Radford et al., 2023) fine-tuned on IPAPack in (Samir et al., 2024).

(Lo)WhIPA[5] is a collection of Whisper-base and -large-v2 models (Radford et al., 2023) which were fine-tuned for STIPA in Suchardt et al. (2025). Aside from parameter size (74M vs. 1550M), the models differ in the utilized fine-tuning strategy (fully vs. Parameter Efficient Fine-Tuning (PEFT) with Low-Rank Adoption (Hu et al., 2021) (LoRA)) and training data. We use the Base and Large variants trained with: Taguchi et al. (2023)'s, i.e. MultIPA's, data (CV) both fully fine-tuned (WhiPA) and LoRA-fine-tuned (LoWhIPA) variants; Arabic Speech Corpus with IPA transcriptions (LoWhIA ASC), and the combination of the CV, ASC, and THCHS-30 data with IPA transcriptions (LoWhIPA Comb.). The model trained on only THCHS-30 had underperformed in the transfer setting and is thus excluded here. While computationally more expensive, (Lo)WhIPA models had excelled on seen language performance compared to MultIPA. Cross-lingual PFER performance was often, yet not always, in favor of MultIPA. A direct comparison of ZIPA and (Lo)WhIPA models has not been presented so far.

Notably, the ZIPA models focus on broad phonetic transcriptions (Zhu et al., 2025): G2P tools were employed to create the IPA transcriptions in their IPAPack++ training data, and all but the 15 most frequent diacritics were removed. Although permissible for the broad transcription of some languages, this limits the model's applicability to languages where the excluded diacritics are contrastive in meaning, e.g., lexical tones in Mandarin. In contrast, regarding the creation of MultIPA, and later (Lo)WhIPA, core training data, Taguchi et al. (2023) characterize the phonemic property of G2P tools as a drawback and control their selection, as well as creation, of conversion tools for phonetic accuracy. Of the LoWhIPA models, which are said to aim for narrow phonetic transcriptions (Suchardt et al., 2025), some further included training data from ASC and THCHS-30 with transcriptions that are more phonetic in nature, without pruning of diacritics. In all studies, models were shown to adapt well to the languages seen during fine-tuning, but challenges in cross-lingual, and cross-domain, transfer remain.

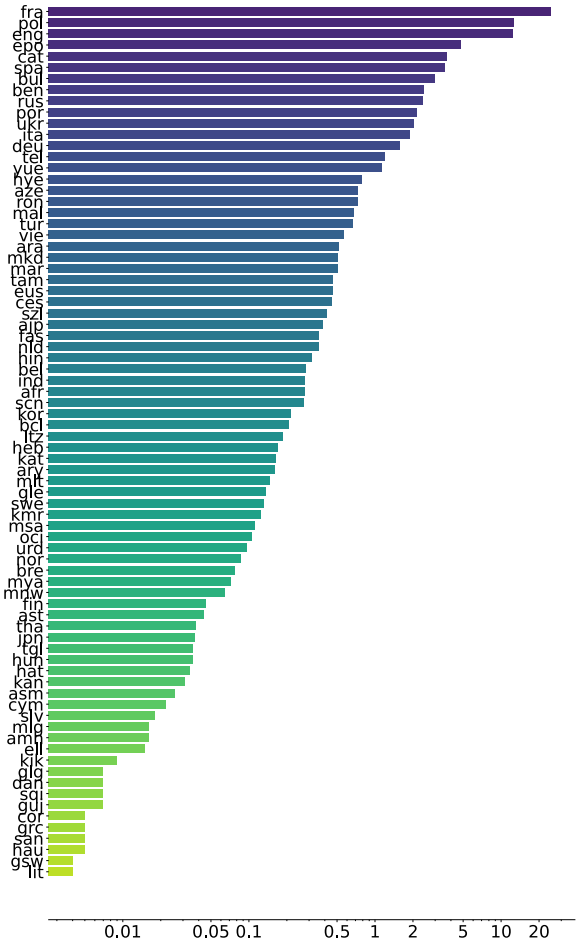

Figure 1: Number of hours per language.

3. WikIPA

The WikIPA dataset represents the integration of two large-scale, community-driven linguistic resources, namely WikiPron and Lingua Libre, to create an open, multilingual benchmark for automatic IPA transcription. While WikiPron provides standardized IPA pronunciation data extracted from Wiktionary, Lingua Libre contributes spoken recordings and detailed speaker metadata. By linking these complementary sources, WikIPA unifies textual and acoustic representations of pronunciation across 78 languages, offering both broad (phonemic) and narrow (phonetic) transcriptions where

[4] https://github.com/lingjzhu/zipa
[5] https://github.com/jshrdt/whipa

Lexical Item	IPA (Broad)	IPA (Narrow)	IPA (Broad) Dialect	IPA (Narrow) Dialect	Speaker	Language
strugarka	[s t r u g a r k a]	—	[None]	—	Poemat	pol
пътуване	[p ɒ t u v ɒ n ɛ]	—	[None]	—	Kiril kovachev	bul
domingo	[d o m i ŋ g o, d o m i ŋ g o]	[d̪ o m ĩ ŋ g o, d̪ o m ĩ ŋ g o]	[ca, la]	[ca, la]	Eavqwiki	spa
gceist	[j ɛ ʃ tʲ]	—	[None]	—	Ériugena	gle
pomme	[p ɔ m]	[pʰ ɔ̃ m]	[None]	[None]	Santamarcanda	fra
inverno	[i n v ɛ r n o]	—	[None]	—	Iopensa	ita
ĩuliteraĵo	[tʃ i u l i t e r a ʒ o]	—	[None]	—	Lepticed7	epo
Götz von Berlichingen	[g ə ts f ʊ n b e r l ɪ ç i ŋ ə n]	—	[None]	—	Brannock	deu
evening	[iː v n ɪ ŋ, iː v ə n ɪ ŋ, i v n ɪ ŋ, iː v n ɪ ŋ, iː v ə n ɪ ŋ]	—	[uk, uk, us, us, us]	—	Nattes à chat	eng

Table 1: Sample entries from the WikIPA dataset. Each row represents a lexical item annotated with its broad and narrow IPA transcriptions, and dialectal variants from the WikiPron database. Although not displayed here, every entry in the dataset also includes the corresponding audio recording and the speaker metadata extracted from Lingua Libre.

Statistic	Train	Test	Total
Number of examples	231,755	57,939	**289,694**
Audio mean (s)	1.15	1.15	**1.15**
Audio min–max (s)	0.29–5.12	0.38–3.7	**0.29–5.12**
Languages	78	78	**78**
Speakers	928	682	**962**

Table 2: WikIPA statistics.

available.

WikiPron is an open-source tool and database for extracting pronunciation data from Wiktionary[6]. It automatically mines word–pronunciation pairs written in IPA across hundreds of languages. The dataset contains approximately 1.7 million pronunciations from 165 languages, each represented as a sequence of graphemes and phones. WikiPron provides either phonemic or phonetic transcriptions depending on the data available in Wiktionary; while some languages include both forms, many contain only one type of transcription. In WikiPron, phonemic transcriptions are also referred to as broad and phonetic transcriptions as narrow. Here, a broad (phonemic) transcription represents only the contrastive sounds, or phonemes, that distinguish words in a language, abstracting away from finer phonetic details. In contrast, the narrow (phonetic) transcription captures a more precise articulatory or acoustic realization of sounds, including allophonic variation and small pronunciation differences. Thus, broad transcriptions are more general and language-specific, while narrow transcriptions are more detailed and speaker- or dialect-specific. In the context of dictionaries in general, and Wiktionary in particular, however, narrow transcriptions should not be interpreted as speaker-specific phonetic records. Rather, they reflect the level of detail provided by Wiktionary contributors for a given language variety, typically a standardized or commonly accepted realization, rather than an instrumentally measured or fully context-dependent articulation.

Lingua Libre is a collaborative, open-access tool and media library developed under the Wikimedia umbrella for recording spoken and signed realizations of words, phrases, and sentences in multiple languages. Contributors use a mass-recording workflow: the interface displays a list of lexical items, which speakers read aloud one by one. The software automatically detects silence between utterances and advances through the list, producing clean, well-trimmed, and consistently named audio clips. These recordings are automatically uploaded to Wikimedia Commons.

In terms of data, Lingua Libre primarily collects audio files linked to lexical entries, i.e. individual words or phrases. Each file is tagged with metadata such as the language, the speaker, and the lexical item recorded. Additional metadata include details about the speaker, such as their place of residence and the languages they speak, along with their proficiency levels.

WikIPA To create the WikIPA dataset, we linked the textual pronunciation data extracted from WikiPron with the corresponding audio recordings from Lingua Libre (samples given in Table 1). All available broad (phonemic) and narrow (phonetic) transcriptions from WikiPron were matched to Lingua Libre entries with identical lexical forms. For each lexical item, multiple transcriptions may be

[6] https://www.wiktionary.org/

present, reflecting alternative pronunciations (e.g., in Italian, the broad transcription of *AIDS* can be either /ajdiɛsse/ or /ajdʒ/). Each transcription is also associated with its dialectal specification when provided in Wiktionary (e.g., *Received Pronunciation* or *General American* for English). When audio data are available, the dataset additionally includes speaker metadata as provided by Lingua Libre. Thus, the WikIPA integration combines WikiPron's standardized IPA transcriptions with Lingua Libre's authentic recordings, providing a unified multilingual resource that connects text-based phonetic representations with empirical speech data. To ensure sufficient coverage, we retained only languages with at least 10 examples. For evaluation purposes, the dataset was further divided into training and test splits, with the test set comprising 20% of the total examples, and the sampling for the test set was stratified by language. Detailed statistics are reported in Table 2, while the total number of hours per language across the full dataset is illustrated in Figure 1.

4. Evaluation

In this section, we introduce the first evaluation of WikIPA using state-of-the-art STIPA models, providing a detailed description of the data, metrics, and results.[7]

4.1. Data

The evaluation is conducted solely on the WikIPA test set and includes only languages with at least 50 test examples, ensuring that each language has a sufficiently large sample size for reliable comparison. For each lexical item, multiple broad or narrow transcriptions may be available, reflecting variation in pronunciation or differences across dialects. While it would be feasible to associate dialect-specific transcriptions with speaker metadata in WikIPA (such as place of residence or proficiency level), this linkage is omitted in the present evaluation for the sake of simplicity and is left for future investigation. In this evaluation, if a lexical item includes both broad and narrow transcriptions, both are retained. However, only one transcription of each type (broad or narrow) is sampled per lexical item.

4.2. Metrics

In this work, we use the revised implementation of the PER and PFER metrics proposed by Suchardt et al. (2025), which addresses key shortcomings of the version proposed by Taguchi et al. (2023).

[7]The code used for the evaluation is available on Github.

Phone Error Rate (PER) is a common metric for assessing phone recognition performance. It is based on the Phone Edit Distance (PED), a phone-level adaptation of the Levenshtein distance. Each phone, including those with diacritics, is treated as a single unit, and all errors are weighted equally, without considering phonetic similarity.

Phonetic Feature Error Rate (PFER) provides a more fine-grained evaluation by comparing phones through their phonetic features. Using PanPhon (Mortensen et al., 2016), phones are represented as 24-dimensional feature vectors with values in -1, 0, +1. PFER is computed as a normalized partial Hamming edit distance, where a full feature mismatch contributes a cost of 1/24, mismatches involving undefined features contribute 1/48, and insertions or deletions incur a cost of 1.

4.3. Results

The results are summarized in the heatmaps shown in Figure 2 and Figure 3, which report respectively the **Phone Error Rate (PER)** and the **Phonetic Feature Error Rate (PFER)** for *broad* and *narrow* transcriptions across the 30 evaluation languages.

Broad Evaluation As visible in the left panels of both figures, PER-performance on broad (phonemic) transcriptions is substantially better than on narrow transcriptions. This reflects the reduced phonological complexity and absence of fine-grained diacritics in the broad setting. Among the evaluated systems, the ZIPA family and particularly the ZIPA-T-SMALL and ZIPA-CR-NS-SMALL variants consistently achieves the lowest mean error rates across languages. This is reflected by the extended areas of yellow/green tones in the left heatmap of Figure 2 for these models. The corresponding PFER patterns in Figure 3 confirm that this advantage holds even when evaluation is performed at the level of articulatory features.

By contrast, the LoWhIPA and WhIPA base models perform considerably worse. Their broad PER values often exceed 60–70 %, particularly for languages such as French. Even WhIPA large configurations do not close the gap with ZIPA entirely, suggesting that architectural efficiency and training diversity, rather than parameter count, are the primary drivers of performance in STIPA. LoWhIPA Large CV performs similar to the MultIPA baseline, as both were trained on only seven orthographically transparent languages, achieving intermediate results. In broad transcription, their mean PER hovers around 56–57 %, several points higher than ZIPA but clearly below the WhIPA models. The difference is especially visible in Italian and French, where ZIPA maintains lower errors.

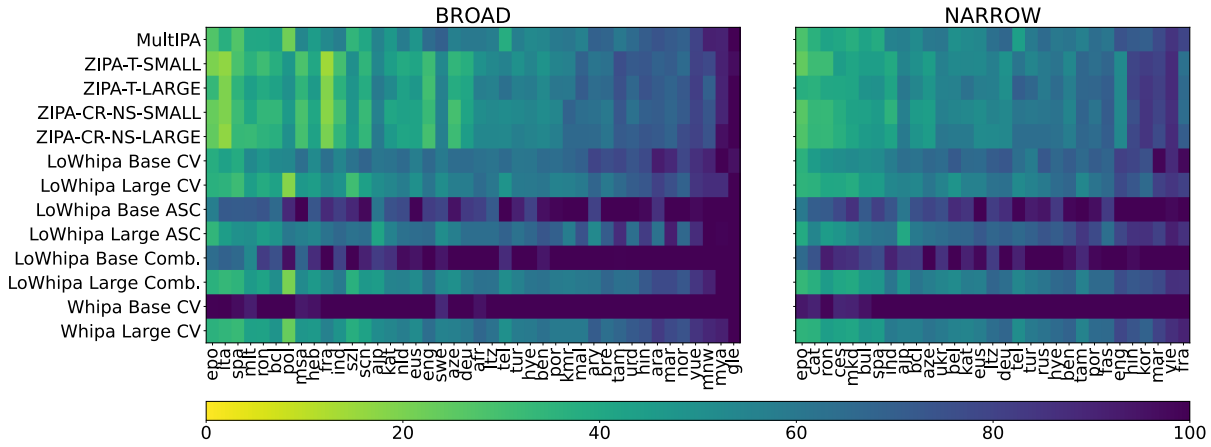

Figure 2: **PER heatmaps.** Each cell represents the average PER for a language–model pair on broad (left) and narrow (right) IPA transcriptions. Cooler colors indicate higher error.

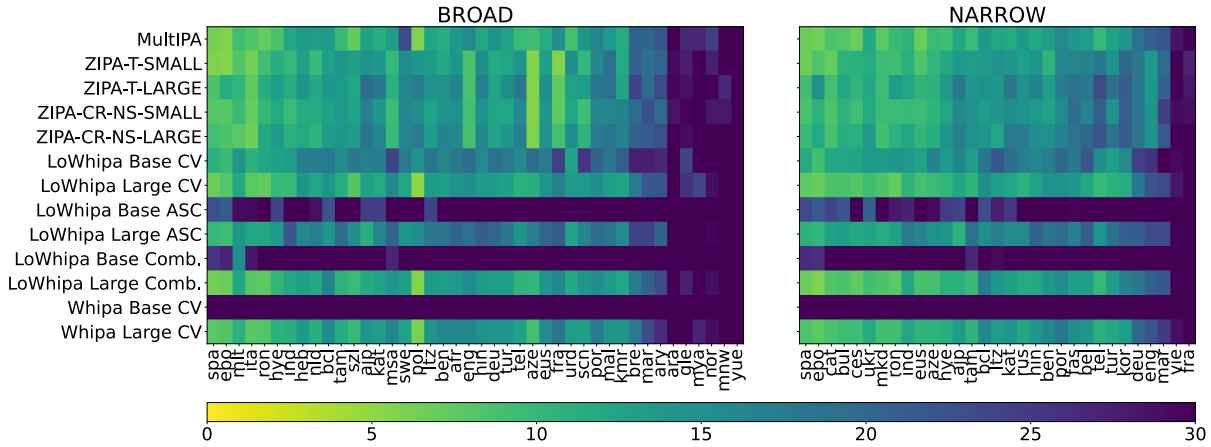

Figure 3: **PFER heatmaps.** Each cell represents the average PFER for a language–model pair on broad (left) and narrow (right) IPA transcriptions. Cooler colors indicate higher error.

However, the lower error rates of ZIPA models must be considered within the context of their STIPA training data which comprised 88 languages and considerably more hours. Among the languages in Figure 2, ZIPA models were directly trained for 30/42 broad PER languages – such as Italian and French – and for 24/31 narrow PER languages. In contrast, MultIPA and (Lo)WhIPA CV models saw only 3 of the broad (mlt, pol, tam) and 1 of the narrow (tam) languages.

Narrow Evaluation When evaluation extends to narrow transcriptions, shown in the right panels of both figures, the error rates rise substantially, as expected from the inclusion of diacritics and secondary articulations. Nonetheless, the relative ranking of models according to PER remains stable. ZIPA-T-SMALL continues to yield the lowest PER overall, closely followed by ZIPA-CR-NS-SMALL and ZIPA-T-LARGE. The narrow PER heatmap (Figure 2, right) shows these models maintaining yellow-green colors across the grid. The WhIPA and LoWhIPA base models continue to exhibit predominantly blue colors. Performance of MultIPA and (Lo)WhIPA Large models is also fairly consistent.

In absolute terms, the ZIPA variants achieve mean narrow PER values around 55 %, representing an improvement of roughly 10 percentage points over WhIPA Base (65–70 %) and 2–4 points over MultIPA and (Lo)WhIPA Large CV/Comb. models. In the narrow PFER heatmap (Figure 3, right), on the other hand, ZIPA-T-SMALL remains below 15 % on average, but is surpassed by MultIPA (12.9 %) and LoWhIPA Large CV (13.0 %), while base variants exceed 40 %.

Effect of Transcription Granularity Languages exhibiting the steepest increase from broad to narrow evaluation are those for which narrow transcription introduces extensive subphonemic detail. To quantify this effect, we computed, for each model

Model	Mean ΔPER	Mean ΔPFER
ZIPA-CR-NS-LARGE	9.40	2.35
ZIPA-CR-NS-SMALL	9.39	2.22
ZIPA-T-SMALL	9.28	2.07
ZIPA-T-LARGE	9.12	2.24
LoWhIPA Base CV	5.76	1.72
MultIPA	5.54	1.29
WhIPA Large CV	5.53	1.24
LoWhIPA Large Comb.	5.52	1.02
LoWhIPA Large CV	5.21	0.92
LoWhIPA Large ASC	4.75	0.58
LoWhIPA Base ASC	3.70	−0.07
WhIPA Base CV	3.68	1.04
LoWhIPA Base Comb.	2.79	−0.07

Table 3: Mean difference (Δ = narrow − broad) in PER and PFER across models. Positive values indicate higher error rates under narrow transcription.

Language	Mean ΔPER	Mean ΔPFER
fra	37.82	23.14
eng	17.60	7.11
spa	12.45	0.68
bcl	10.05	0.63
eus	6.78	−4.71
deu	6.75	3.87
kat	6.33	0.29
por	5.44	−1.33
hye	5.19	0.65
tur	4.55	1.59
ben	4.14	−0.07
tel	3.58	1.50
mar	3.48	0.49
ind	2.49	−1.24
ltz	2.46	−0.52
hin	1.96	−1.11
ajp	1.56	−0.81
aze	1.15	−4.11
ron	0.92	0.45
epo	0.87	0.29
tam	−6.92	−0.03

Table 4: Mean difference (Δ = narrow − broad) in PER and PFER across languages. Positive values indicate higher error rates under narrow transcription. Only languages available in both settings are included.

and language represented in both annotation types, the difference in error rate between narrow and broad transcription (Δ = narrow − broad) for both PER and PFER. Languages without a narrow transcription set were excluded to ensure comparability. The computation was carried out over all models. All models show positive Δ values (Table 3) for PER, confirming that narrow transcription systematically increases model error. The effect is most pronounced for the ZIPA variants, whose mean ΔPER ranges from 9.1 to 9.4 and mean ΔPFER from 2.0 to 2.3, while MultIPA and WhIPA models display smaller differences (PER $\approx$ 4–6, PFER $<$2). This indicates that despite their overall superior accuracy, ZIPA models are more sensitive to the additional phonetic detail introduced in the narrow setting. At the language level, the highest ΔPER values (Table 4) are observed for French (+37.8) and English (+17.6), followed by Spanish (+12.4), German (+6.8), and Basque (+6.7). In contrast, languages such as Bengali, Marathi, Hindi, and Tamil show minimal or slightly negative differences, suggesting that their narrow and broad transcriptions are more closely aligned. Overall, these results confirm that transcription granularity has a measurable and language-dependent impact on STIPA performance, with larger degradations in languages where narrow transcriptions encode fine-grained phonetic variation.

Cross-metric Consistency A close inspection of Figures 2 and 3 reveals a strong correlation between PER and PFER. Improvements in symbolic accuracy translate proportionally into reductions in feature-based error. The PFER heatmaps display less variance across typologically similar languages, suggesting that the feature representation provides a more reliable estimate of cross-linguistic phonetic similarity than purely symbolic PER.

Interestingly, the ZIPA-CR-NS variants show nearly identical PFER values to their ZIPA-T counterparts despite slightly higher PERs, implying that their internal representations capture articulatory proximity even when the predicted phone sequence diverges. This points to the potential benefits of feature-aware objectives or articulatory-space alignment for future work.

Model Scale and Architecture The comparison between small and large model variants offers further insight into architectural efficiency. In both figures, the small ZIPA models often outperform their larger counterparts, indicating that the Zipformer architecture benefits from compact parameterization and mitigates overfitting. This supports the findings of Zhu et al. (2025), which show that smaller models tend to generalize better. Conversely, (Lo)WhIPA's performance improves with scale, as larger models achieve lower PER and PFER.

4.4. Error Analysis

To complement the quantitative evaluation, we conducted a qualitative analysis of model outputs across three languages: Italian, German, and English. For each language, we sampled 100 prediction–reference pairs from the *broad* transcription subset. Samples were drawn proportionally to each model's overall PFER error distribution, ensuring that the subset reflected the same performance hierarchy observed in the full evaluation set, while

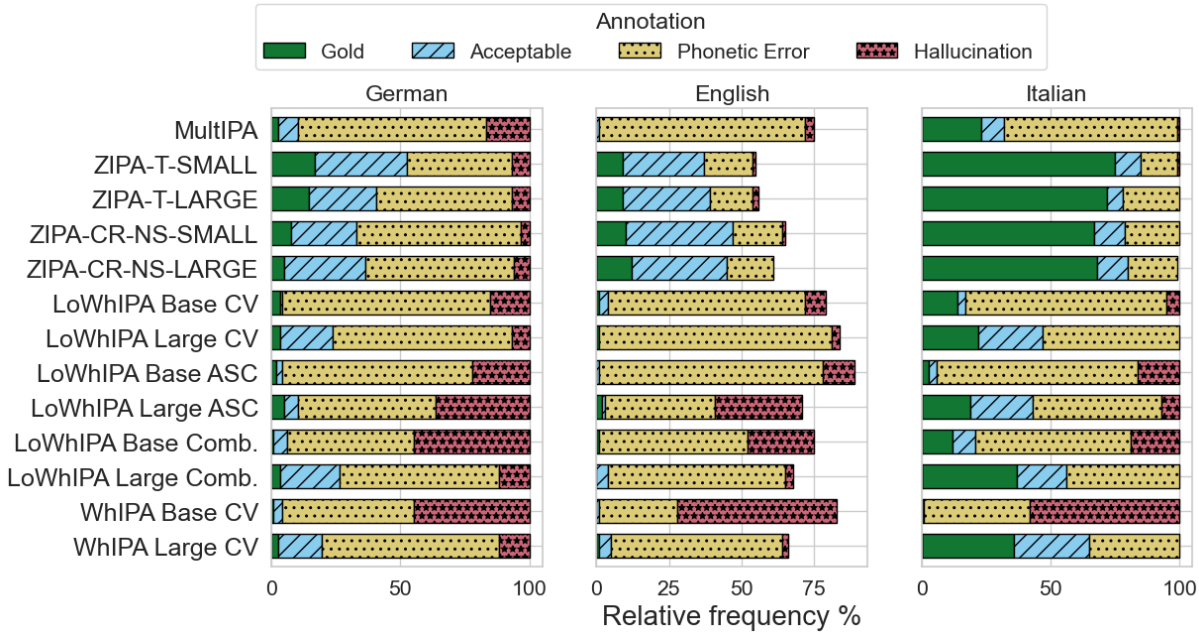

Figure 4: Relative frequency of annotation categories across models for German, English, and Italian. Each stacked bar shows the proportion of samples labeled as *Gold* (correct), *Acceptable* (minor variant), *Phonetic Error*, or *Hallucination*.

preventing annotators from identifying which system produced each prediction.

Each sampled prediction was manually inspected and annotated using a four-category error typology:

1. **Transcription Mismatches/Acceptable Variation:** Instances where multiple plausible IPA transcriptions exist for the same word or variant. For example, the Italian *r* may appear as [r] or [ɐrre]; similarly, English vowels often show dialectal variation (e.g., [ɑ] vs. [ɒ] in *cot*). These mismatches reflect ambiguity in the reference rather than model failure.

2. **Phonetic Error:** Substitutions or distortions involving phonetically similar segments (e.g., [s] vs. [ʃ] , [r] vs. [ɾ]), suggesting confusion in fine-grained acoustic distinctions rather than complete model failure.

3. **Model Hallucinations:** Cases in which the model produced an output entirely unrelated to the input speech, often including orthographic tokens.

4. **Errors in Lingua Libre recordings:** As Lingua Libre is a crowdsourced corpus, some recordings were misaligned or mislabeled. Typical errors involved speakers recording multiple words consecutively or uttering unintended material.

Inter-annotator reliability was evaluated on the English subset; two annotators independently labeled the 100 sampled instances per model. The highest agreement was obtained for the WhIPA Base CV ($\kappa = 0.66$) and LoWhIPA Base ASC ($\kappa = 0.63$) models, where the distribution of error categories was highly consistent between annotators. In both cases, categories 2 (phonetic errors) and 4 (alternative transcriptions) were labeled almost identically, while categories 0 and 1 occurred infrequently and showed minimal disagreement. Intermediate values were observed for the ZIPA models ($\kappa = 0.26$–0.42), indicating fair-to-moderate agreement, while the lowest scores were found for MultIPA ($\kappa = 0.16$) and WhIPA Large CV ($\kappa = 0.18$). Models with low agreement still exhibited similar distributions of error categories overall, discrepancies lie in the disagreement over the exact nature of error vs. plausibly acceptable transcriptions. On average, the micro-averaged agreement across all models was $\kappa = 0.38$, and the macro-average reached $\kappa = 0.51$, corresponding to moderate agreement between the annotators.

Beyond agreement analysis, Figure 4 summarizes the relative frequency of annotation categories across models for German, English, and Italian. Only the samples for which the two annotators fully agreed are included for English. Category 4 (recording or corpus errors) is not represented. This error category was established following observations of the data, but were not present in the annotated samples. The remaining categories are shown in the following order: category 0 (*Gold*), category 1 (*Acceptable* or plausible alternative tran-

scription), category 2 (*Phonetic Error*) and category 3 (*Hallucination*). The overall distributions reveal consistent cross-linguistic patterns: most model outputs fall within the *Phonetic Error* category, followed by *Gold* or *Acceptable* transcriptions for ZIPA models, while hallucinations are more common in monolingual and Base (Lo)WhIPA models. English and German exhibit a higher proportion of phonetic errors, reflecting greater phonetic and dialectal variability as well as higher degrees of vowel reduction and coarticulation. Italian, by contrast, shows a larger share of *Gold* and *Acceptable* predictions, likely due to its more stable segmental articulation and clearer vowel–consonant contrasts, which facilitate a closer acoustic match with the reference transcriptions. These distributions complement the quantitative evaluation: Although model accuracy varies – as expected since ZIPA models were trained on all three annotated languages – the qualitative nature of errors remains systematically patterned across languages.

5. Conclusion

We introduced WikIPA, an open multilingual benchmark that links human-curated IPA transcriptions from WikiPron with speech and speaker metadata from Lingua Libre, enabling systematic evaluation of STIPA systems across 78 languages and 289k audio–IPA pairs. By supporting both broad (phonemic) and narrow (phonetic) transcriptions, WikIPA connects symbolic phonetic representations to empirical speech at scale. Our evaluation of state-of-the-art models shows ZIPA variants, especially ZIPA-T-SMALL and ZIPA-CR-NS-SMALL most frequently achieve the lowest mean error rates on both broad and narrow settings, outperforming (Lo)WhIPA and MultIPA in most setting but narrow PFER evaluation. Narrow transcription predictably raises error rates for many systems, with the largest degradations in languages where narrow transcriptions encode fine-grained subphonemic detail (e.g., French and English). The results also reflect the distinct modeling foci of current systems: ZIPA variants were primarily optimized for broad (phonemic) transcription, MultIPA exhibits limited adaptation to narrow forms, and (Lo)WhIPA models explicitly target narrow phonetic detail. Consequently, differences between broad and narrow evaluations are larger for ZIPA and smaller for (Lo)WhIPA, consistent with their respective training objectives and data sources. A targeted error analysis across Italian, German, and English indicates that most residual model errors are minor phonetic confusions rather than outright failures; hallucinations are comparatively rarer and concentrated in some (Lo)WhIPA base configurations. These findings suggest that the central challenge for STIPA is modeling fine-grained articulatory variation.

6. Bibliographical References

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.

Martin J Ball and Joan Rahilly. 2014. *Phonetics: The science of speech*. Routledge.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021*, pages 2426–2430.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

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

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.

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

Qiantong Xu, Alexei Baevski, and Michael Auli. 2021. Simple and effective zero-shot cross-lingual phoneme recognition.

Xuanru Zhou, Jiachen Lian, Cheol Jun Cho, Tejas Prabhune, Shuhe Li, William Li, Rodrigo Ortiz, Zoe Ezzes, Jet Vonk, Brittany Morin, et al. 2025. Towards accurate phonetic error detection through phoneme similarity modeling. In *Proc. Interspeech 2025*, pages 4738–4742.

7. Language Resource References

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.

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.

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.

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.

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.

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.

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.

Jacob Lee Suchardt, Hana El-Shazli, and Pierluigi Cassotti. 2025. Towards Language-Agnostic STIPA: Universal Phonetic Transcription to Support Language Documentation at Scale. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China. Association for Computational Linguistics.

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.

Dong Wang and Xuewei Zhang. 2015. Thchs-30 : A free chinese speech corpus.

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.

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