A Comparative Survey and Evaluation of Language-Agnostic Speech-to-IPA Transcription Systems

Allan Valim Ribeiro da Fonseca
Applied Computer Science Department
University of Applied Sciences Fulda
Fulda, Germany
allan.valim-ribeiro-da-fonseca@informatik.hs-fulda.de

***Abstract*—Transcribing spoken language is challenging: the auditory signal is complex, and a transcriber often revises a segment repeatedly before choosing a symbol, yielding only an approximation. This bottleneck has long restricted linguistic fieldwork (often manually transcribed). Recent self-supervised speech models can transcribe audio into the International Phonetic Alphabet (IPA). We surveyed seven such approaches and evaluated the publicly available ones against a shared benchmark. This addresses a gap in the literature since each system used a different evaluation protocol and reported different metrics. Each is scored using both Phone Error Rate (PER) and Phonological Feature Error Rate (PFER). Removing narrow detail (diacritics, length, tone) reduces PER by 9 to 22 points while barely affecting PFER, showing this penalty comes from fine detail in the references, not recognition error. Finally, training-data quality appears to matter more than training-data size.**

***Index Terms*—universal phone recognition, phonetic transcription, multilingual speech recognition, language documentation, phone error rate, comparative evaluation.**

I. Introduction

The International Phonetic Alphabet (IPA) was developed by the International Phonetic Association [1], led by Paul Passy, with the intent of facilitating and formalizing the transcription of spoken languages. First published in 1888 and revised several times since, it assigns a distinct character to most sounds the human vocal tract can produce, making differences between sounds clearer in writing.

IPA is often associated only with documenting languages at the brink of extinction, but its uses are broader: linguists also compare how different dialects pronounce the same words, or how a language evolved over time. Distinguishing the sounds used in pronunciation matters because spelling changes far less often than dialects do.

The use of IPA is not limited to linguists, since it is also relevant, e.g., for the diagnosis of certain medical conditions. Given that transcribing one minute of audio takes roughly one hour of tedious and repetitive work, influenced by factors such as the transcriber's experience or native language, the result is inherently subjective. Thus, recent advances in automating this task are of great interest. In Computational Linguistics, Allosaurus (2020) [2] was an influential early system for universal IPA phone recognition.

Across the papers investigated in this survey, we observe not only a lack of common benchmarks, but also an inconsistency in their reported error rates. This work seeks to bridge that gap by using the same evaluation set to compare all these implementations.

In this paper we: (1) survey 7 systems from 2020–2025, (2) evaluate the public models on a shared gold benchmark (human-audited references), then a silver one (G2P-predicted) used by some papers, (3) show that reference granularity, rather than recognition error, accounts for a substantial share of the narrow-vs-broad scoring difference, and (4) find that fine-tuning with a small high quality dataset remains competitive with large-scale approaches.

The rest of this paper covers background (Section II), the survey (Section III), our methodology (Section IV), results (Section V), and conclusions (Section VI).

II. Background

The following list explains terms referenced throughout this text.

IPA: International Phonetic Alphabet, a collection of symbols that list all the currently understood sounds producible from a human mouth. **Phone**: Each physical speech sound represented by the IPA, e.g. [p] in *pin*. **Phoneme**: The contrasting sounds understood as different by a language, e.g. /p/ vs. /b/ in *pat*/*bat*. **Allophone**: Each phone variant that represents the same phoneme, e.g. [p^h] in *pin* vs. [p] in *spin*.

Phonetic Features: 21 ways phones are produced (tongue position, air exhalation type, jaw height, mouth-rounding, etc.) **Diacritics**: Small marks added to a base IPA symbol to indicate finer, mostly segmental detail. They mark: articulatory changes, like vowel nasalization; subtle sound differences, like aspiration; some also mark suprasegmentals such as tone and stress, like tones in Mandarin. **Broad**: Simpler transcription. Ignores or significantly reduces phonetic features / diacritics ("fold" them to the most notable phone), e.g. Hindi word [putʃ]. **Narrow**: Stricter transcription. Considers most of the

diacritics, e.g. [pũːt͡ʃʰ], carrying nasalization, length, aspiration and an affricate tie bar.

G2P: Grapheme-to-Phoneme, it means to transliterate normal text into IPA. **PanPhon** [3]: relates an IPA symbol to its phonetic features. Used to calculate PFER. **Epitran** [4]: G2P tool used by most papers to generate training data from audio-text-pair datasets.

PER: Phone Error Rate. Indicates how close to the *exact* reference-transcription each model's output is. **PFER**: Phonological Feature Error Rate. Calculates how many of the 21 attributes each symbol in a transcription has and scores how close it is to the reference. **PTER**: Phone Token Error Rate. Like PER, but each diacritic and suprasegmental mark counts as its own token rather than being attached to a phone. **AER**: Attribute Error Rate. Allophant's metric; scores errors over articulatory attributes rather than whole phones, conceptually close to PFER.

VoxAngeles (VA) [5]: Gold benchmark. Corpus of human-checked phonetic transcriptions and phone-level alignments of the UCLA Phonetics Lab Archive; contains data from 95 languages across 21 language families, plus unaudited automatic alignments for another 11 languages. Recordings are individual words. Files were removed where the audio did not match the transcription, contained only noise, or was cut off mid-word; inconsistent suprasegmental notations (tone, syllable boundaries, stress) were also removed during auditing. **FLEURS** [6]: Silver benchmark. Google's corpus of 102 languages. Audio of read-speech with textual transcription. **Common Voice (CV)** [7]: Mozilla's crowd-sourced corpus of volunteers reading sentences. **PHOIBLE** [8]: Database of phoneme inventories for over 2,000 languages; used to build language-specific phone sets. **BABEL**: IARPA corpus of conversational telephone speech across many languages. **MLS** [9]: Multilingual LibriSpeech; read audiobook speech in several languages. **THCHS-30** [10]: Mandarin read-speech corpus. **ASC** [11]: Arabic Speech Corpus; read South Levantine (Damascan) Arabic. **IPAPack** [12]: Multilingual speech corpus with IPA transcriptions; IPAPack++ audits and expands upon it. **DoReCo** [13]: Documentation Reference Corpus; naturalistic fieldwork speech from 50 mostly endangered languages, transcribed by linguists.

III. Survey

The seven papers we survey below are related to the following models, in chronological order: Allosaurus [2], Wav2Vec2Phoneme [14], MultIPA [15], Allophant [16], ZIPA [17], WhIPA [18], and POWSM [19]. They were chosen for executing speech-to-IPA, having public weights, reporting PER/PFER, and being published after 2020 in peer-reviewed venues.

The first distinction we can make between the implementations is whether they were trained from scratch or fine-tuned from a pre-trained model, and what their size is (total parameters). Allosaurus, ZIPA, and POWSM belong to the first group: Allosaurus did not report its total size (estimated around 11M), ZIPA trained models with both 64M and 300M parameters, and POWSM trained with 350M. Except for WhIPA, with sizes of 74M (Base) and 1550M (Large), Wav2Vec2Phoneme, MultIPA, and Allophant used about 300M. In other words, except for two, all papers trained at least one 300M configuration, making their results more directly comparable.

Regarding the first group's architectures, Allosaurus trained its own acoustic model using BiLSTM plus an extra allophone layer per language trained, ZIPA used a Zipformer [20] with CR-CTC transducer, and POWSM used AED encoder-decoder (ESPnet [21]) because it wanted to multi-task across PR (phone recognition), ASR (Automatic Speech Recognition), G2P and P2G (Phoneme-to-Grapheme) instead of specializing in PR alone. As for the fine-tuned group, aside from WhIPA using Whisper [22] as its pre-trained model, the others have a similar base: both MultIPA and Wav2Vec2Phoneme used wav2vec2 [23] XLSR-53 [24] and CTC encoder, and Allophant used XLS-R with multi-task on attribute classifiers. This is what the pre-trained models offered: XLSR-53 counted with 53 languages for training, while XLS-R had 128, and Whisper had 99.

For their entire models, Allosaurus uses 11 languages, based on the PHOIBLE database, while ZIPA created IPAPack++, ending up with over 17,000 hours of audio from 88 languages, plus another 11,000 hours of pseudo-labeled audio produced by noisy-student training. POWSM also trained on IPAPack++. As for the fine-tuning details, Wav2Vec2Phoneme used extra 52,000 hours of data with 42 languages from CV, 19 from BABEL, and 6 from MLS. Since its base was already bigger, Allophant used 4,628 hours with 34 non-tonal languages from CV. Finally, WhIPA finds MultIPA the most robust of the earlier systems under cross-lingual transfer: it shows that a small set of high-quality curated data (1,000 utterances per language) can produce results that remain competitive to this day; more specifically, they used only 9 hours of data (7 languages) for fine-tuning. WhIPA followed suit with 7 languages from CV, then THCHS-30 and ASC, choosing 1000 utterances for each language.

Manual transcription at that scale is impractical, so all implementations used artificially generated transcriptions, made with different G2P tools (most commonly eSpeak NG, Epitran [4], and CharsiuG2P [25]). MultIPA, ZIPA, and WhIPA later challenged the premise that large data is needed at all. To circumvent the loss of accuracy from G2P tools, that often "correct" output into their grammatically standardized version, all papers curated the transcriptions.

WhIPA had a strictly narrow transcription. ZIPA normalized them, only kept the 15 most common diacritics, and if there were more than three, it was reduced to one. Even after simplification, ZIPA still is considered narrow. On the broad transcription variants, POWSM removes all break and length diacritics, Allophant ignores all tone, Allosaurus and Wav2Vec2Phoneme have only phonemic transcriptions. MultIPA "removed" tones and suprasegmental diacritics by choosing languages in which transcription and normal text have little difference (e.g., unlike English).

Allosaurus tested against 11 seen languages (present in

the training data) and two unseen (Inuktitut and Tusom). Wav2Vec2Phoneme had a CV benchmark of 13 languages, 6 from MLS (one used as zero-shot), and 6 seen from BABEL with 2 unseen (Cantonese, Lao). Allophant tested on 34 seen and 84 unseen languages taken from the UCLA Phonetic Corpus (which was later refined into VA, which we use for our benchmark). WhIPA went with 7 seen and 4 unseen languages (emulating MultIPA), while also testing against Mandarin (THCHS-30), Levantine (ASC mono), and using 95 from VA for zero-shot evaluation.

Although ZIPA only evaluated on unseen data, with 95 tongues from VA, and 45 from DoReCo, it was one of the two papers that investigated how transcriptions differ sociolinguistically (i.e., dialects), using two English-exclusive corpora for that. POWSM also investigates 5 English corpora, while also evaluating on 11 seen languages from IPAPack++, 45 and 95 unseen from DoReCo and VA, respectively, plus Tusom, like Allosaurus. MultIPA stands out for the quality of its evaluation rather than its scale: apart from being tested on its 7 seen languages, their 4 unseen (Luganda, Upper Sorbian, Hakha Chin, Tatar) were transcribed by two trained human annotators.

Based on available models, we recalculated results using PER and PFER. Even though PER is strict, a model that scores low under it has proven its efficacy. PFER grades more leniently, so it flatters outputs that are far from ideal: a wrong transcription can still score well when its IPA symbols are ”close enough” (see Table I). WhIPA further notes that PFER handles length discrepancies poorly, producing counter-intuitive scores. Allosaurus reported PER, while MultIPA introduced PFER and reported PER for comparison. Wav2Vec2Phoneme had PER and PTER. Allophant showed PER and created the AER metric. WhIPA was the second paper to use both PER and PFER (third, if AER is counted as a PFER variant). Finally, ZIPA and POWSM reported only PFER, the more forgiving of the two metrics.

Side note: One curiosity that comes to light when reading the authors' section of these papers, is that researchers from ByteDance, Meta, Google, and Intel Labs were listed on them (Allosaurus, Wav2Vec2Phoneme, MultIPA, and Allophant, respectively), indicating an interest of ”big tech” in the area.

IV. Methodology

Across the seven papers, we evaluated ten configurations (accounting for multi-variant setups) to replicate their results on our benchmarks. The three extra versions came from ZIPA and WhIPA. ZIPA had the large version, comparable with the size observed in other implementations (300M), as well as the small one (64M), which was inspired by MultIPA's relatively compact fine-tuning layer. In comparison, WhIPA has the base Whisper model of 74M with their fine-tuning, plus another ”base” version fine-tuned with the LoRA technique [26] (version known as LoWhIPA). WhIPA trained a model with 1550M parameters to verify whether a bigger model would achieve better results. As for the other models, we used versions publicly available, or the ones that, according to their original paper, showed results different enough to be relevant.

All ten configurations were scored on the complete gold benchmark: 5,416 utterances covering 95 languages in 21 language families, and on the full silver benchmark. Note that `whipa_comb` is the combined-data *LoRA* adapter (LoWhIPA), so the LoRA axis is represented in our results at base size.

Our gold benchmark follows the three most recent papers, which also used it, most notably the WhIPA paper, which our tests expand upon. For that, we use the VoxAngeles corpus [5]: the closest available corpus to the audited narrow transcriptions we wanted to test against.

Because we tested models on out-of-scope languages, we designed a silver benchmark that more closely emulates the datasets the models trained on. We chose the FLEURS dataset [6] over CV because it offers an API call for each audio clip, accompanied by a normal transcription. Then, we used the G2P Epitran [4] to generate an evaluation set. The surveyed papers used the same method, except that they curated the output. After evaluation, our results were similar to the ones reported.

These references are *silver*: Epitran output on the FLEURS transcripts, uncurated and phonemic rather than phonetic. The silver benchmark is therefore a benchmark-design demonstration, showing what changes when the reference style changes, and not a ranking of the systems.

The evaluation required setting up different environments to avoid dependency conflicts. Few approaches offered terminal commands or repository, sharing instead weights and code snippets. After installation, we made scripts to automate the file preparation and test execution[1]. It was necessary to convert VA files to 16 kHz to match what the models were trained on.

The script: wrapped all available code for running the models in helper functions; called them for each file on the evaluation set folder; wrote the outputs in TSV files; used PanPhon [3] to traverse these lists; compared each result to its reference transcription[2]; and calculated the error rates. Later, we summarized the results per file[3], per family[4] and per model[5].

V. Results

Table I shows each model configuration's transcription hypothesis for the Icelandic word ”bið” (to abide). The audio clip had [b iː ð] as transcription for its pronunciation; for reference, its phonemic transcription (standard pronunciation) is [ˈp ɪː ð]. The table shows the exact outputs.

Notice that each implementation in Table I has two PER and PFER results: narrow (left) and broad (right). Narrow is the result against the entire unaltered VoxAngeles reference. For

[1]Available at https://github.com/allan-valin/phonetic-eval/, along with instructions on how to reproduce the tests manually or automatically.
[2]References.
[3]Per file results.
[4]Summary per family.
[5]Per model results.

the broad result, we remove diacritics, length and tone marks from both the reference and the hypothesis and rescore. Note how the rows whose errors are genuine phone substitutions barely move, while a hypothesis that differed only by narrow detail drops to zero. To generate this table, we did an evaluation round on this single audio file, then saved the outputs.

The results in Table II served as calibration: if our numbers were similar to the ones published, we took this as evidence that our calculations and metrics were faithful. The differences observed come from scoring the whole of VoxAngeles rather than the language subsets each paper chose.

Table III represents the results from our gold benchmark. The first two PER and PFER columns show the reported error rates in the papers. Since Allophant did not report PFER, but rather AER, we merged the columns, owing to their similarity. ZIPA and POWSM also evaluated on VA, and on its full unseen set, yet report PFER an order of magnitude below ours. The gap therefore does not come from an easier language subset; it comes from what their references contain, both having removed diacritic and length detail from their pipelines, and from reporting the more forgiving of the two metrics.

Our recomputed error rates are higher across the board, since we test each system on out-of-scope data to check whether it is truly language-agnostic. Allosaurus reported 25% PER, while our broad testing showed 82.6%; its PFER, however, stayed at 32.5% even in these unfavorable circumstances. POWSM reported only PFER, at 2.6%, but our testing revealed 15.1% PFER and 45.7% PER.

Table IV shows the silver benchmark with our calculated PER and PFER results, based on G2P transcriptions of 6 languages, 5 families, 120 FLEURS utterances. Comparing these results with Table III, PER falls by a quarter to a half depending on the configuration, halving for the three systems trained on IPAPack++, but still stays above the reported results. This drop cannot be attributed to reference style alone: the silver set also differs in audio domain (isolated words vs. read sentences) and language set (95 mostly unseen vs. 6 high-resource), and the three IPAPack++-trained systems match the broad G2P reference style they were trained on.

Lastly, comparing the same configurations on different sizes, ZIPA's 300M variant did not show lower error rates than its 64M. WhIPA is the opposite case: going from the whisper-base backbone to whisper-large-v2, on the same narrow Common Voice training data, cuts narrow PER by 14.3 points on the gold benchmark (83.7 for whipa, 69.4 for whipa_large) and by 24.3 on the silver one (62.9 vs 38.6). The balanced 720-utterance subset gives a delta of 14.0, so the effect is not an artifact of which utterances were sampled. Expanding the training corpora to include Mandarin and Arabic helped as well.

In WhIPA's paper, LoRA variants (LoWhIPA) showed better performance than full fine-tuning at both sizes. Our `whipa_comb` configuration is one of these adapters, so the technique is represented here at base size; the large LoRA variant was left out because each whisper-large-v2 pass costs about ten hours on our hardware, as much as every other configuration combined.

A. Limitations

The silver benchmark differs from the gold one on three axes simultaneously: reference style (Epitran G2P vs. human audit), audio domain (connected sentences vs. isolated words), and language set. Thus, it demonstrates the effect of evaluation choices rather than isolating any single cause, and is not a ranking. It is also far smaller (120 utterances, 6 languages, 5 families) than the gold set (5,416 utterances, 95 languages, 21 families), so its scores are indicative only. The silver references are uncurated, phonemic G2P output, not phonetic transcriptions. Allophant's reported figure in Table III is AER rather than PFER, so that cell is not strictly like-for-like. Finally, whipa_large was evaluated only without LoRA, as each large pass costs roughly ten hours, leaving the LoRA axis covered at base size alone.

VI. Conclusion

Machine-learning models for speech-to-IPA have come a long way, each approaching the problem differently. That is exactly why comparing them against the same baseline matters: future work can better judge where to start. We found a noticeable distinction between reported results and what we could reproduce when applying the models to their target use-case: a language-agnostic system transcribing any language it meets.

A substantial part of that distinction is not recognition error. Reducing both sides of the comparison to broad phones recovers between 9 and 22 PER points while leaving PFER nearly untouched; the rest of the reported-vs-observed gap comes from evaluating on out-of-scope languages. The systems are mostly hearing the right phone and being penalized in full for narrow detail their training data never contained. This is a scoring-time analysis and therefore a different operation from ZIPA's, which removes diacritics from training and evaluation data and retrains; the two results do not contradict each other.

G2P-generated training sets were a common thread among the surveyed papers and were often acknowledged as something to improve upon. We recommend that future work train and evaluate on narrowly transcribed utterances produced by professional annotators; while prohibitive for a small team, a community-wide effort—building on existing curated fieldwork archives, with due regard for the access rights of the source communities—could yield better training data than what was observed.

Given the subjectivity of the task, MultIPA [15] points out that there is a 20% difference between transcriptions from human annotators. Thus, we would argue that decreasing the PER to this number should be the community's goal. While PFER is relevant to evaluate the suprasegmental deviations in transcriptions, Table I reveals that the most usable outputs are the ones with both error rates low. If the results are "close enough", but still require the linguist to swap all symbols, it would probably require less work to transcribe it from scratch.

References

[1] International Phonetic Association, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press, 1999. doi: 10.1017/9780511807954.

[2] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black, and F. Metze, "Universal phone recognition with a multilingual allophone system," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8249–8253. doi: 10.1109/ICASSP40776.2020.9054362.

[3] D. R. Mortensen, P. Littell, A. Bharadwaj, K. Goyal, C. Dyer, and L. Levin, "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*, Y. Matsumoto and R. Prasad, Eds. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 3475–3484. [Online]. Available: https://aclanthology.org/C16-1328/

[4] D. R. Mortensen, S. Dalmia, and P. Littell, "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), May 2018. [Online]. Available: https://aclanthology.org/L18-1429/

[5] E. Chodroff, B. Pažon, A. Baker, and S. Moran, "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)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 12 724–12 733. [Online]. Available: https://aclanthology.org/2024.lrec-main.1114/

[6] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "FLEURS: Few-shot learning evaluation of universal representations of speech," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2023. doi: 10.1109/SLT54892.2023.10023141.

[7] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*. Marseille, France: European Language Resources Association (ELRA), 2020, pp. 4218–4222. [Online]. Available: https://aclanthology.org/2020.lrec-1.520/

[8] S. Moran and D. McCloy, "PHOIBLE 2.0," Max Planck Institute for the Science of Human History, Jena, 2019. doi: 10.5281/zenodo.2593234, available: https://phoible.org.

[9] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," in *Interspeech 2020*, 2020. doi: 10.21437/Interspeech.2020-2826.

[10] D. Wang and X. Zhang, "THCHS-30: A free chinese speech corpus," *arXiv preprint arXiv:1512.01882*, 2015. doi: 10.48550/arXiv.1512.01882.

[11] N. Halabi, "Modern standard Arabic phonetics for speech synthesis," Ph.D. dissertation, University of Southampton, 2016. [Online]. Available: https://eprints.soton.ac.uk/409695/

[12] J. Zhu, C. Yang, F. Samir, and J. Islam, "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)*. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 750–772. doi: 10.18653/v1/2024.naacl-long.43. [Online]. Available: https://aclanthology.org/2024.naacl-long.43/

[13] L. Paschen, F. Delafontaine, C. Draxler, S. Fuchs, M. Stave, and F. Seifart, "Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo)," in *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*. Marseille, France: European Language Resources Association (ELRA), 2020, pp. 2657–2666. [Online]. Available: https://aclanthology.org/2020.lrec-1.324/

[14] Q. Xu, A. Baevski, and M. Auli, "Simple and Effective Zero-shot Cross-lingual Phoneme Recognition," in *Interspeech 2022*, 2022, pp. 2113–2117. doi: 10.21437/Interspeech.2022-60.

[15] C. Taguchi, Y. Sakai, P. Haghani, and D. Chiang, "Universal Automatic Phonetic Transcription into the International Phonetic Alphabet," in *Interspeech 2023*, 2023, pp. 2548–2552. doi: 10.21437/Interspeech.2023-2584.

[16] K. Glocker, A. Herygers, and M. Georges, "Allophant: Cross-lingual Phoneme Recognition with Articulatory Attributes," in *Interspeech 2023*, 2023, pp. 2258–2262. doi: 10.21437/Interspeech.2023-772.

[17] J. Zhu, F. Samir, E. Chodroff, and D. R. Mortensen, "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)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 19 568–19 585. doi: 10.18653/v1/2025.acl-long.961. [Online]. Available: https://aclanthology.org/2025.acl-long.961/

[18] J. L. Suchardt, H. El-Shazli, and P. Cassotti, "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, Nov. 2025, pp. 31 423–31 439. doi: 10.18653/v1/2025.emnlp-main.1600. [Online]. Available: https://aclanthology.org/2025.emnlp-main.1600/

[19] C.-J. Li, K. Chang, S. Bharadwaj, E. Yeo, K. Choi, J. Zhu, D. R. Mortensen, and S. Watanabe, "POWSM: A phonetic open whisper-style speech foundation model," *arXiv preprint arXiv:2510.24992*, 2025. doi: 10.48550/arXiv.2510.24992.

[20] Z. Yao, L. Guo, X. Yang, W. Kang, F. Kuang, Y. Yang, Z. Jin, L. Lin, and D. Povey, "Zipformer: A faster and better encoder for automatic speech recognition," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2310.11230.

[21] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Interspeech 2018*, 2018. doi: 10.21437/Interspeech.2018-1456.

[22] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 202, 2023, pp. 28 492–28 518. doi: 10.48550/arXiv.2212.04356.

[23] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020. doi: 10.48550/arXiv.2006.11477.

[24] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020. doi: 10.48550/arXiv.2006.13979.

[25] J. Zhu, C. Zhang, and D. R. Mortensen, "ByT5 model for massively multilingual grapheme-to-phoneme conversion," in *Interspeech 2022*, 2022. doi: 10.21437/Interspeech.2022-538.

[26] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2022. doi: 10.48550/arXiv.2106.09685.

Artefacts

The tables below report the full per-configuration results; our previous text discusses only the most notable values.

TABLE I

WORKED EXAMPLE ON ICELANDIC INDOEUROPEAN_ISL_001. A NEAR-CORRECT PHONE DIFFERING ONLY BY A DIACRITIC OR LENGTH MARK. VALUES % (NARROW → BROAD).

Model	Hypothesis	PER	PFER
allosaurus	b̥ i ð	66.7 → 0.0	1.4 → 0.0
wav2vec2phoneme	b̥ iː θ	33.3 → 33.3	1.4 → 1.4
allophant	b ɪ̯ ɪ d̪	66.7 → 66.7	33.3 → 33.3
multipa	piθ	100.0 → 66.7	4.2 → 2.8
zipa (small)	p i t s	133.3 → 100.0	38.2 → 36.8
zipa_large	b i θ	66.7 → 33.3	2.8 → 1.4
powsm	p e iː θ	100.0 → 100.0	33.3 → 33.3
whipa	biz	66.7 → 33.3	4.2 → 2.8
whipa_comb	piːtː	66.7 → 66.7	6.9 → 5.6
whipa_large	piːθ	66.7 → 66.7	2.8 → 2.8

TABLE II

CALIBRATION: CONFIGURATIONS WITH PUBLISHED VOXANGELES NUMBERS, SO OUR COLUMNS *ARE* DIRECTLY COMPARABLE. THE MULTIPA ROW IS THE VOXANGELES SCORE REPORTED FOR IT BY WHIPA, NOT BY ITS OWN PAPER.

Config	New PER	Old PER	New PFER	Old PFER
multipa	62.4	60.1	15.0	15.4
whipa	83.7	87.8	27.1	32.1
whipa_comb	77.8	81.6	19.3	23.0
whipa_large	69.4	70.7	17.9	20.5

TABLE III

EACH SYSTEM AS REPORTED IN ITS OWN PAPER VERSUS OUR RECOMPUTED RESULTS AVERAGED OVER VOXANGELES (95 LANGUAGES, 21 LANGUAGE FAMILIES, 5,416 UTTERANCES). THE REPORTED COLUMNS ARE *NOT* COMPARABLE WITH EACH OTHER: MOST ARE IN-DOMAIN SCORES, WHILE THE WHIPA ROWS ARE THAT PAPER'S OWN ZERO-SHOT VOXANGELES NUMBERS. OUR VA COLUMNS SHOW NARROW → BROAD. ALLOPHANT'S REPORTED "PFER" CELL IS ITS AER, A RELATED BUT NON-IDENTICAL METRIC.

Config	PER	PFER	VA PER	VA PFER
allosaurus	25.0	—	91.7 → 82.6	32.7 → 32.5
wav2vec2phoneme	22.2	—	53.0 → 41.7	15.4 → 13.6
allophant	45.6	19.4	71.0 → 50.6	20.5 → 18.6
multipa	21.0	5.7	62.4 → 50.3	15.0 → 14.1
zipa (small)	—	3.2	64.2 → 50.4	18.0 → 15.9
zipa_large	—	2.7	69.1 → 54.5	20.3 → 18.1
powsm	—	2.6	55.5 → 45.7	16.1 → 15.1
whipa	87.8	32.1	83.7 → 68.8	27.1 → 25.4
whipa_comb	81.6	23.0	77.8 → 55.4	19.3 → 17.1
whipa_large	70.7	20.5	69.4 → 56.1	17.9 → 16.8

TABLE IV

CONTRAST SET: FLEURS SENTENCES SCORED AGAINST BROAD *SILVER* G2P (EPITRAN) REFERENCES, WHICH ARE PREDICTED, NOT AUDITED. THIS MEASURES AGREEMENT WITH A BROAD G2P REFERENCE AND IS A BENCHMARK-DESIGN DEMONSTRATION, NOT A RANKING. VALUES NARROW → BROAD.

Config	PER	PFER
allosaurus	64.2 → 57.2	18.9 → 18.8
wav2vec2phoneme	36.9 → 34.6	10.8 → 10.5
allophant	49.7 → 44.9	16.7 → 16.1
multipa	47.5 → 41.6	13.7 → 13.6
zipa (small)	32.5 → 27.7	7.7 → 7.0
zipa_large	31.5 → 26.0	7.6 → 6.8
powsm	23.7 → 22.3	8.2 → 7.8
whipa	62.9 → 53.9	18.9 → 18.7
whipa_comb	55.2 → 42.6	14.6 → 13.9
whipa_large	38.6 → 34.6	11.3 → 11.2