Transcription of (Un-)seen Languages with Speech-to-IPA Models

Jacob Lee Suchardt
Leipzig University & ScaDS.AI Dresden/Leipzig
jacob.suchardt@uni-leipzig.de

Relevant UniDive working groups: WG1, WG3

1 Introduction

The predominance of English in language technology applications and Natural Language Processing (NLP) (Bender, 2009) has been gaining visibility (Schwartz, 2022), as research strives to become more linguistically diverse and inclusive. Most commonly, a lack of data [for a specific task] (Nigatu et al., 2024) is cited as the reason why some languages are catered to by research and technology, while others (often called "low resource languages") are seemingly left behind. However, for many NLP technologies, there is also an implicit demand that the training data for current, data-hungry models be in the form of standardised, digitised orthographical text. In contrast, many of the world's languages are primarily (or exclusively) oral and may have either no written form at all, or one that is non-standardised. As a consequence, NLP models often handle variation poorly (Bergmanis et al., 2020), aspects of sociolinguistic variation are missed, and, despite being the most common modality for human natural language, the particularities of spoken language are comparatively under-researched.

We add to the recently emerging studies in speech-to-IPA which use Automatic Speech Recognition (ASR) models specifically trained to predict phonetic transcriptions in a language-agnostic format to i) take first steps towards making spoken data tangible across more linguistically diverse and low-resource languages, ii) add greater attention to sociolinguistic detail within languages, and iii) avoid standardisation or imposing orthographies.

2 Approach

ASR has improved vastly over the past decade, even achieving human-like performance in some cases (Radford et al., 2023; Yadav and Sitaram, 2022). However, standard Speech-to-Text applications produce standardised orthography – losing sociolinguistic and paraverbal detail – and training such models is often not a viable option for the aforementioned under-resourced languages. Manual transcription, on the other hand, is highly time-consuming, especially when phonetic transcriptions are desired (Seifart et al., 2018).

Therefore, we aim to predict phonetic transcriptions in the largely universal International Phonetic Alphabet (IPA) from speech directly. Our study contrasts with the few previous end-to-end speech-to-IPA models, which often use Wav2vec2 (Baevski et al., 2020) as the base (Xu et al., 2021; Glocker et al., 2023; Taguchi et al., 2023), by building on the autoregressive encoder-decoder Transformers model Whisper (Radford et al., 2023), whose encoder received highly multilingual pretraining, helping to create language-universal feature encodings (Pratap et al., 2020; Feng et al., 2021; Zhu et al., 2024; Yadav and Sitaram, 2022), We set a novel "ipa" language token and train on all languages conjointly to encourage the model to learn a language-agnostic mapping between phonetic signals and IPA symbols and update only the decoder weights. Since paired speech-IPA data are rare, we adopt the data conversion scripts from Taguchi et al. (2023) and fine-tune on 1,000 samples each from 7 languages of CommonVoice (Ardila et al., 2020). Ideally, if sufficient coverage over all human language sounds could be achieved in the training data, the model should be able to predict phonetic transcriptions on any unseen language.

3 Results

We evaluate using string-level Phone Error Rate (PER), which takes absolute (dis-)agreement of phones as the smallest unit, and Phonetic Feature Error Rate (PFER), which compares phones based on phonetic feature vectors.

CommonVoice Data Table 1 displays model performance on the test sets of the fine-tuning languages (left) and unseen, held-out languages (right) for fine-tuned Whisper models (WhIPA, LoWhIPA fine-tuned with PEFT LoRA) and the Wav2vec2-based MultIPA model by Taguchi et al. (2023) for

	Model	el	fi	hu	ja	mt	pl	ta	M1	cnh	hsb	lg	tt	M2
PER	MultIPA	24.0	36.8	36.4	25.2	**18.9**	29.3	37.6	29.7	**73.5**	**60.4**	76.0	**62.4**	**68.1**
	WhIPA Base	50.4	49.9	42.2	40.9	57.4	68.3	61.6	53.0	89.3	72.7	84.8	92.3	84.8
	WhIPA Large	9.7	7.8	19.6	11.3	20.6	16.3	22.0	15.3	81.6	68.8	73.9	76.3	75.1
	LoWhIPA Base	29.8	46.6	34.8	37.1	45.2	55.5	47.5	42.4	82.5	68.5	80.5	75.7	76.8
	LoWhIPA Large	**8.3**	**7.6**	**17.2**	**10.5**	20.1	**15.5**	**20.1**	**14.2**	81.7	65.8	**72.6**	71.9	73.0
PFER	MultIPA	7.3	9.3	8.9	7.9	**7.0**	10.2	11.0	8.8	**20.1**	**17.6**	**20.6**	**17.3**	**18.9**
	WhIPA Base	12.7	12.8	13.3	21.9	17.9	26.2	19.1	17.7	26.3	21.4	24.0	33.4	26.3
	WhIPA Large	**3.7**	**2.4**	5.4	3.3	7.6	**6.6**	8.0	5.3	24.7	22.9	25.6	24.0	24.3
	LoWhIPA Base	7.8	11.2	11.4	19.5	11.8	17.8	14.3	13.4	23.0	18.4	21.0	21.8	21.1
	LoWhIPA Large	3.9	2.5	**4.7**	**2.9**	7.5	6.9	**6.6**	**5.0**	25.6	20.1	27.8	21.8	23.9

Table 1: Per-language and macro mean (M) PER/PFER scores per model (**bold**: best performance, underlined: second-best).

Model	CV		Unrelated		Overall	
	PER	PFER	PER	PFER	PER	PFER
MultIPA	**46.8**	**9.3**	**64.67**	**16.57**	**60.11**	**15.38**
WhIPA Base	88.7	39.6	87.93	31.60	87.80	32.09
WhIPA Large	50.3	12.4	74.32	21.35	70.66	20.45
LoWhIPA Base	61.6	15.3	76.18	21.95	74.39	21.41
LoWhIPA Large	51.4	14.5	70.32	19.90	66.55	19.08

Table 2: Average PER and PFER scores on VoxAngeles.

comparison. Performance improves along with the model size and with the addition of LoRA fine-tuning. On unseen languages (human-authored gold labels, 25 per language), performances deteriorates, the performance gap between Large and Base models' scores is reduced, and MultIPA excels. Thus, in-domain (language-wise) performance does not guarantee out-of-domain performance, which suggests that large models may overfit to training languages and, in return, emphasises the need to include unseen languages during development to monitor transfer performance. The metrics do not scale linearly: A model that achieves similar PFER scores across two languages may exhibit dissimilar PER scores for the same pair, e.g., WhIPA Base for Greek/Hungarian.

VoxAngeles We test the validity of the previous results on VoxAngeles (Chodroff et al., 2024), a highly diverse dataset spanning 5k utterances across 314 languages with human-authored phonetic gold transcriptions. The results in Table 2 are split into a CV-group (languages also found in the speech-to-IPA fine-tuning data), an Unrelated group, and an Overall macro average. Models perform better on familiar languages (CV-group) than on unfamiliar ones, but produce higher error rates on this novel data distribution. MultIPA shows the lowest performance loss in the CV-group, underlining its adaptability to varying data sources.

4 Application to Sanna

Finally, we deploy the models to interview recordings with speakers of Sanna (also: Cypriot Maronite Arabic) (Roth, 1979; El-Shazli, 2024a,b), an endangered Arabic vernacular spoken by fewer than 2,000 individuals in Kormakiti, Cyprus. Error rates, compared to an expert phoneticians gold transcriptions, are stable (c.a. 20%), and in favour of a LoWhiPA Large model using only typologically related Greek, Maltese, and Arabic for fine-tuning. Moreover, analyses revealed that erroneous phones typically belong to the same broad phonetic class (Loweimi et al., 2023) as the target phone, deviate only in minor phonetic details (diacritics), and often display plausible phonological processes (e.g., assimilation). This is promising for their legibility and thus their viability for use by annotators.

5 Conclusion

We have presented a systematic evaluation of speech-to-IPA models to support phonetic annotations for spoken languages. Our findings suggest that autoregressive Whisper models are particularly suited for in-domain settings and potentially few-shot settings, while alternative models such as MultIPA may excel in true zero-shot settings.

Relevance for UniDive UniDive WG1 T1.5 reports show spoken treebanks on the rise, albeit with unique challenges, and that <25% have a phonetically informed transcription. Speech-to-IPA models could help contributors provide spoken language treebanks with more complete phonetic transcription coverage. For WG3 T3.1, robust speech-to-IPA pipelines could also be the first step towards creating more data resources for under-resourced languages, later aiding the development of multilingual tools, such as morphosyntactic tagging.

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.

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.

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

Toms Bergmanis, Artūrs Stafanovičs, and Mārcis Pinnis. 2020. Robust Neural Machine Translation: Modeling Orthographic and Interpunctual Variation. In *Human Language Technologies – The Baltic Perspective - Proceedings of the Ninth International Conference Baltic HLT 2020*, pages 80–86, Kaunas, Lithuania. IOS Press.

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.

Siyuan Feng, Piotr Żelasko, Laureano Moro-Velázquez, Ali Abavisani, Mark Hasegawa-Johnson, Odette Scharenborg, and Najim Dehak. 2021. How phonotactics affect multilingual and zero-shot asr performance. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7238–7242.

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

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.

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The zeno's paradox of 'low-resource' languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics.

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.

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.

Lane Schwartz. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 724–731, Dublin, Ireland. Association for Computational Linguistics.

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

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.

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

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.

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.