An (unhelpful) guide to selecting the right ASR architecture for your under-resourced language

Robbie Jimerson
RIT
rcj2772@rit.edu

Zoey Liu
University of Florida
liu.ying@ufl.edu

Emily Prud'hommeaux
Boston College
prudhome@bc.edu

Abstract

Advances in deep neural models for automatic speech recognition (ASR) have yielded dramatic improvements in ASR quality for resource-rich languages, with English ASR now achieving word error rates comparable to that of human transcribers. The vast majority of the world's languages, however, lack the quantity of data necessary to approach this level of accuracy. In this paper we use four of the most popular ASR toolkits to train ASR models for eleven languages with limited ASR training resources: eleven widely spoken languages of Africa, Asia, and South America, one endangered language of Central America, and three critically endangered languages of North America. We find that no single architecture consistently outperforms any other. These differences in performance so far do not appear to be related to any particular feature of the datasets or characteristics of the languages. These findings have important implications for future research in ASR for under-resourced languages. ASR systems for languages with abundant existing media and available speakers may derive the most benefit simply by collecting large amounts of additional acoustic and textual training data. Communities using ASR to support endangered language documentation efforts, who cannot easily collect more data, might instead focus on exploring multiple architectures and hyperparameterizations to optimize performance within the constraints of their available data and resources.

1 Introduction

The majority of significant academic and industry research on automatic speech recognition (ASR) (Povey et al., 2011; Hinton et al., 2012; Amodei et al., 2016; Watanabe et al., 2018; Baevski et al., 2020) has been evaluated on a small set of English language datasets (Panayotov et al., 2015; Godfrey et al., 1992). Word error rates (WER) for English ASR now approach those of human transcriptionists (Baevski et al., 2020; Radford et al., 2022), and speakers of English can now reliably use ASR for text entry when using mobile devices. This level of accuracy, however, is attainable only for the handful of the world's 7000 languages that, like English, have abundant training resources.

Most of the world's languages, even ones spoken by tens of millions of speakers, currently lack datasets prepared specifically for training ASR models. The datasets that do exist are typically much smaller than English ASR datasets that have been available for decades, with no more than a few dozen hours of acoustic training data. As the Common Voice project (Ardila et al., 2020) has shown, collecting large amounts of data for widely spoken languages is possible, but using this kind of platform is likely to be impractical for the roughly 40% of the world's languages that are endangered (Eberhard et al., 2022). A similar percentage of languages – again, even many that are widely spoken – lack an established writing system, which presents other obstacles to building large ASR corpora.

Fortunately, existing methods for training accurate ASR models for English and other high-resource languages can be adapted to low-resource settings. Some toolkits include recipes for smaller datasets that require the training of fewer parameters. Other approaches rely on fine-tuning acoustic models pre-trained on massive multilingual speech datasets. Most recent work using these approaches, however, does not compare the performance of multiple competitive architectures across multiple diverse small ASR datasets. Thus, while we have access to transformative technology that can be harnessed to build reasonable models for languages with limited resources, we do not know which of the popular architectures is "better" or whether features of a particular dataset or language might make one architecture more suitable than another.

In this paper we explore four different popular ASR architectures, three of which are currently considered state of the art, that can be used even in

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1008–1016
July 9-14, 2023 ©2023 Association for Computational Linguistics

Language		HH:MM	# Speakers	# LM tokens	Audio quality	Audio source
Bemba	train	17:17	8	96K	variable	read speech
	test	02:00	2			
Wolof	train	16:49	14	600K	high	read speech
	test	00:55	2			
Swahili	train	10:00	N/A	3M	variable	read speech and broadcast news
	test	01:45	N/A			
Seneca	train	09:57	11	76K	variable	fieldwork
	test	02:04	11			
Fongbe	train	07:35	25	990K	high	read speech
	test	01:45	4			
Iban	train	07:00	17	200K	high	broadcast news
	test	01:00	6			
Hupa	train	06:06	1	41K	variable	fieldwork
	test	01:31	1			
Oneida	train	03:23	7	18K	variable	fieldwork
	test	00:51	4			
Quechua	train	03:00	N/A	8.1K	variable	conversations
	test	00:45	N/A			
Bribri	train	00:29	N/A	4K	variable	fieldwork
	test	00:11	N/A			
Guarani	train	00:19	N/A	1.2K	variable	read speech
	test	00:07	N/A			

Table 1: Characteristics of the eleven datasets. The datasets for Bemba, Fongbe, Wolof, and Iban were partitioned by holding one or more speakers out to serve as test data. Information about the exact number of speakers for Swahili, Quechua, Bribri, and Guarani, and about whether any speaker is represented in both the test set and training set was not explicitly provided in the dataset or the accompanying paper. Train/test partitioning for the Hupa, Seneca, and Oneida datasets was done randomly; some or all speakers are represented in both the training and test sets. We note that for some of these datasets, very long and very short utterances had to be removed due to the training constraints of one or more of the ASR architectures. For this reason, the audio times and token counts reported here may differ from those reported in the associated papers or those that would be derived directly from the unfiltered data.

low-resource settings: a hybrid DNN (Veselỳ et al., 2013); two approaches for fine-tuning from a multilingual pre-trained acoustic model (Conneau et al., 2020; Radford et al., 2022); and an end-to-end approach designed specifically for small datasets (Shi et al., 2021). We train models for eleven datasets for under-resourced languages, which are diverse in their linguistic properties, mechanisms for collection, relative sizes, and recording quality.

We find that no single approach to training ASR models in low-resource settings consistently outperforms any other, with the most outdated method turning out to be the most accurate surprisingly often. While unsatisfying in some ways, these results can help guide ASR researchers and language community members to select the architecture that is most compatible with their objectives and that can be feasibly supported with their available financial and personnel resources. For widely spoken languages, where the goal of developing an ASR system is likely to be to support a voice-based app or a personal digital assistant, the best use of financial resources might be to collect large amounts of additional data in order to take advantage of state-of-the-art high-resource architectures. Linguists and members of endangered language communities hoping to use ASR to document and preserve their language cannot easily gather more data, and thus might see more benefit from carefully experimenting with multiple architectures to identify the approach that provides the best results for their particular language or existing dataset.

2 Related Work

Although most of the notable advances in ASR have focused on English and a few other languages with abundant data, there has been substantial inter-

Language Name	Language Family	Language Status	Morphological Properties	Tonal (Y/N)	Number of Phones
Bemba	Niger-Congo	education (4)	agglutinative	Y	27
Wolof	Niger-Congo	wider communication (3)	agglutinative	N	41
Swahili	Niger-Congo	national (1)	agglutinative	N	37
Seneca	Iroquoian	endangered (8a)	polysynthetic	N	23
Fongbe	Niger-Congo	wider communication (3)	isolating	Y	33
Iban	Austronesian	wider communication (3)	agglutinative	N	25
Hupa	Eyak-Athabaskan	endangered (8b)	polysynthetic	N	44
Oneida	Iroquoian	endangered (8a)	polysynthetic	N	17
Quechua	Quechuan	wider communication (3)	agglutinative	N	33
Bribri	Chibchan	endangered (6b)	agglutinative	Y	32
Guarani	Tupian	national (1)	polysynthetic	N	31

Table 2: Linguistic properties of the eleven languages explored here. Language status is the EGIDS reported in Ethnologue (Eberhard et al., 2022). Phone counts are taken from Ethnologue, Glottolog (Hammarström et al., 2022), or the paper reporting the dataset. All eleven languages are written primarily using the Roman alphabet with diacritics to indicate features such as nasality, vowel length, and tone.

est in ASR for languages with minimal training resources for quite some time (Besacier et al., 2014). Much of the work from the 2010s focused on the languages of the IARPA Babel project (Thomas et al., 2013; Miao et al., 2013; Cui et al., 2014; Grézl et al., 2014). Research initiated with the Babel datasets on methods of transfer learning and data augmentation in low-resource settings has continued apace (Khare et al., 2021; Vanderreydt et al., 2022; Guillaume et al., 2022b). With the success of the Kaldi toolkit, researchers began to collect and freely distribute their own Kaldi-ready datasets for under-resourced and endangered languages, several of which are explored in this paper (Gauthier et al., 2016; Laleye et al., 2016; Gelas et al., 2012; Juan et al., 2015; Pulugundla et al., 2018). More recent work has explored training monolingual end-to-end models with substantially larger datasets than those used here (Shi et al., 2021), as well as transfer learning and fine-tuning from pretrained multilingual (Guillaume et al., 2022a; Sikasote and Anastasopoulos, 2022) or English models (Thai et al., 2020).

3 Datasets

Five of the datasets explored here are freely available datasets built by researchers, sometimes in collaboration with speech communities, specifically for training ASR models for widely spoken but under-resourced languages of the global South: Bemba (Sikasote and Anastasopoulos, 2022), Fongbe (Laleye et al., 2016), Wolof (Gauthier et al., 2016), Swahili (Gelas et al., 2012), and Iban (Juan et al., 2014, 2015). Three datasets (Quechua, Bribri, Guarani) were created from existing recordings for the 2022 AmericasNLP Workshop Shared Task [1]. The remaining datasets for three endangered languages of North America (Hupa, Oneida, and Seneca) were created using existing linguistic and community fieldwork recordings available to the authors through the affiliation of one of the authors with one of these communities and the generosity of the community elders.

While nearly any recorded speech can be transcribed and used to train an ASR system, a common approach for building a new ASR dataset is to ask speakers of the language to read aloud provided texts, which obviates the laborious task of transcription. With this strategy, speakers are often recorded in a studio or similarly controlled environment, resulting in more consistent recording quality. Alternatively, datasets can be created from existing audio data such as radio broadcasts or linguistic fieldwork recordings. Such recordings are often already transcribed but need to be segmented and

[1] http://turing.iimas.unam.mx/americasnlp/2022_st.html

time-aligned with the transcripts, which must often be done by hand. Table 1 provides details about these sorts of characteristics of the datasets, as well as information about the quantity of the training data for the acoustic and language models.

Information about the linguistic characteristics of the eleven languages is provided in Table 2. Seven of these languages are widely spoken by millions of people, and some have institutional or government recognition; one is endangered with around 7,000 speakers; and three are critically endangered with very few (perhaps only one, in the case of Hupa) first-language speakers and no more than a hundred second language learners. A diverse set of morphological, phonological, and phonetic features and properties are represented among these languages, and we note that they are all quite different typologically from most high-resource languages, including not only English and Chinese but also the major European languages.

4 ASR Architectures

The goal of this work is to explore whether any one of several popular and state-of-the-art ASR architectures is especially well suited for building models with small amounts of training data. We train models on the the eleven datasets described in Section 3 using four different architectures:

- A hybrid DNN (Veselỳ et al., 2013) implemented within the Kaldi toolkit (Povey et al., 2011), following Karel's DNN recipe[2] which uses a variety of feature optimizations including RMB pretraining, frame cross-entropy training, and MBR sequence-discriminative training. Decoding was performed with a trigram language model.

- A transducer-based end-to-end model for small datasets within ESPnet2 (Watanabe et al., 2018), following the recipe for Yoloxochitl Mixtec (Shi et al., 2021).

- Fine-tuning from a multilingual acoustic model using Wav2Vec2 XLSR-53 (Conneau et al., 2020), decoding both with and without a trigram language model and using the parameterizations specified in the Hugging Face Wav2Vec XLSR-53 tutorial.[3]

- Fine-tuning from the medium multilingual acoustic model with Whisper (Radford et al., 2022), using the parameterizations specified in the Hugging Face Whisper tutorial.[4]

Training and testing were carried out on a university high-performance computing cluster. Training times ranged between 2 and 24 hours depending on the architecture and dataset.

5 Results

Figure 1 shows the word error rates (WER) for four of the five approaches (Kaldi DNN, Wav2Vec XLSR with and without a language model (LM), and Whisper) when trained and tested on each of the eleven datasets. Note that prior baselines reported in the papers associated with the datasets for Wolof, Swahili, Fongbe, Hupa, and Iban, using non-s.o.t.a. architectures, and Bemba, using a slightly different configuration of Wav2Vec XLSR, are lower than the best reported architecture here. No prior WER results have been reported for the Oneida, Quechua, Bribri, and Guarani datasets.

We observe a large variation in WER across languages, which should not be surprising given the great variability in the quantity of training data, the type and audio quality of data collected, and the linguistic features of these languages. Datasets of less than 3 hours had consistently high WERs, but across the other datasets, there does not appear to be a clear relationship between amount of audio training data and WER. Though not shown in Figure 1, ESPnet yielded the worst performance by far for all languages, with only Wolof, the second largest dataset, achieving a WER below 65%. Again, this is not surprising given that this ESPnet recipe (Shi et al., 2021) was proposed for a much larger 60-hour indigenous language dataset.

More interestingly, we see no consistent ranking of the remaining four approaches across the eleven datasets. Using an LM during decoding with Wav2Vec XLSR always yields some improvement in WER over not using an LM, but the differences are often quite small. Notably, Swahili, which has the largest LM, sees only a tiny reduction in WER when that LM is used during decoding. The Kaldi hybrid DNN, despite being outdated, outperforms more than one of its state of the art rivals for Seneca, Fongbe, Iban, and Quechua. Whisper is dramatically better than other models for Wolof and Hupa,

[2] https://kaldi-asr.org/doc/dnn1.html

[3] https://huggingface.co/blog/fine-tune-xlsr-wav2vec2

[4] https://huggingface.co/blog/fine-tune-whisper

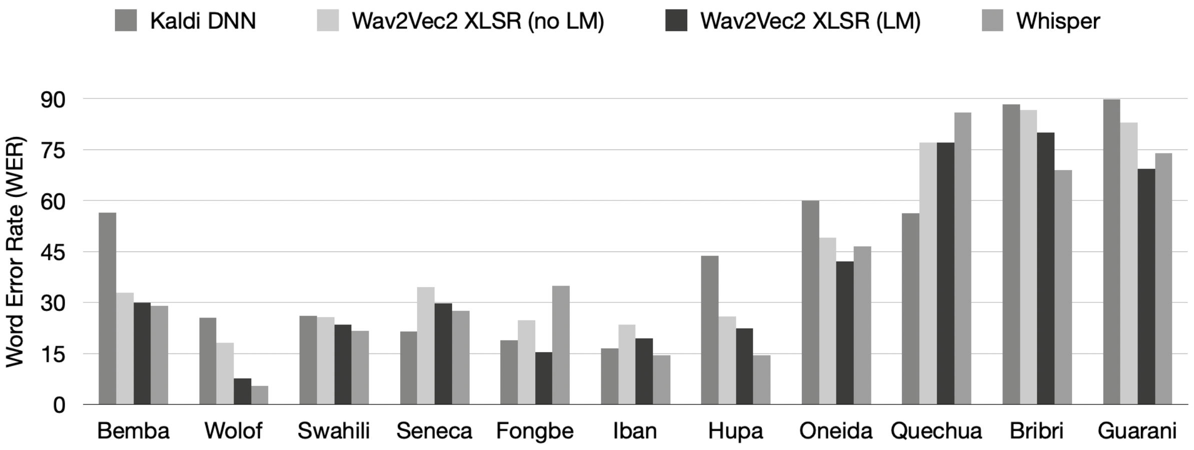

Figure 1: Word error rates (WER) for each dataset under each architecture. A lower WER indicates fewer errors and higher accuracy. There is no clear winner among the other models, with even the Kaldi hybrid DNN often outperforming more recent state of the art fine-tuning approaches. Where WER has been previously reported, one or all of these models outperform the reported baselines in the published papers for these datasets.

but substantially worse for Fongbe and Quechua. Though closely related and typologically similar, Seneca and Oneida show very different patterns, as do Fongbe and Wolof, two related languages with datasets recorded under similar conditions. The WER for Swahili is relatively stable across architectures, while WER is quite variable for Wolof, Hupa, Fongbe, and Oneida.

The rankings do not appear to be related to the method of speech collection (read vs. spontaneous) or the consistency of audio quality. In addition, whether or not a language is tonal, like Bemba, Fongbe, and Bribri, does not appear to predict the relative rankings of the four architectures.

We do note, however, two potential patterns, which merit further investigation with a larger set of languages. First, Fongbe, the only language of the eleven with isolating morphology (i.e., limited affixation) is one of only two languages where Whisper yielded the highest WER of the four systems. Second, the three languages with the largest phonesets, Wolof, Swahili, and Hupa, yielded the same relative ranking, with Whisper performing the best and Kaldi the worst. Although there is certainly not enough information here to draw conclusions, it is plausible that the design of a particular training architecture or the content of the pretrained models could render a system more appropriate for a language with a particular linguistic property.

6 Conclusions

Under-resourced language communities, whether large or small, need to know how to invest their limited resources when developing an ASR system for their language. Our findings suggest, unfortunately, that there are no obvious or simple guidelines to follow. Our future work will expand the set of languages explored here in order to establish connections between expected model performance and linguistic features and dataset characteristics. We also plan to explore the impact of language model size and domain on ASR accuracy and the relationship between language model and morphology.

Limitations

One limitation of this work is that we have included results for only eleven languages. Training ASR models, even on small datasets, requires significant computing and financial resources. Second, there are not that many freely available and well prepared ASR datasets that are readily compatible with all four ASR architectures. We sought to select a diverse set of languages and datasets with varying features in order to provide, we hope, a reasonable snapshot of how the state of the art performs in low-resource settings.

Ethics Statement

The Hupa, Oneida, and Seneca datasets were recorded with the approval of participating universities' IRBs and with the enthusiastic cooperation of the elders and other linguistic consultants. The datasets for the remaining languages were downloaded from public Web pages. The Bribri dataset, like those of other endangered languages, was created using linguistic fieldwork recordings. Of the others, some were collected by recruiting participants to read text (Wolof, Fongbe, Bemba, Guarani); others consist of transcribed radio and television broadcasts (Iban, Quechua); and the Swahili dataset includes both types of data. While the participants who provided recordings by reading text presumably gave consent for their voices to be used for ASR research, it is unlikely that speakers recorded in the course of a radio or television broadcast provided consent explicitly for their voices to be used in an ASR dataset. We expect, however, given that members of the speech community participated in these data collection projects, that ethical concerns were carefully considered.

Acknowledgements

We are grateful for the continued support from the Hupa and Seneca indigenous communities. We would like to especially thank Mrs. Verdena Parker, of the Hoopa Valley Tribe, and Mrs. Sandy Dowdy, of the Seneca Nation of Indians, for their generous and valuable input and support. This material is based upon work supported by the National Science Foundation under Grant #2127309 to the Computing Research Association and Grant #1761562. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Computing Research Association.

References

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, et al. 2016. Deep speech 2 : End-to-end speech recognition in english and mandarin. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 173–182.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222.

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.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Xiaodong Cui, Brian Kingsbury, Jia Cui, Bhuvana Ramabhadran, Andrew Rosenberg, Mohammad Sadegh Rasooli, Owen Rambow, Nizar Habash, and Vaibhava Goel. 2014. Improving deep neural network acoustic modeling for audio corpus indexing under the IARPA Babel program. In *Fifteenth Annual Conference of the International Speech Communication Association*.

David M Eberhard, Gary F. Simons, and Charles D. Fennig. 2022. *Ethnologue: Languages of the World. Twenty-fifth edition*. SIL International.

Elodie Gauthier, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Elingui. 2016. Collecting resources in sub-Saharan African languages for automatic speech recognition: a case study of Wolof. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3863–3867, Portorož, Slovenia. European Language Resources Association (ELRA).

Hadrien Gelas, Laurent Besacier, and Francois Pellegrino. 2012. Developments of Swahili resources for an automatic speech recognition system. In *SLTU - Workshop on Spoken Language Technologies for Under-Resourced Languages*, Cape-Town, Afrique Du Sud.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Frantisek Grézl, Martin Karafiát, and Karel Vesely. 2014. Adaptation of multilingual stacked bottle-neck neural network structure for new language. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 7654–7658. IEEE.

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. 2022a. Fine-tuning pre-trained models for automatic speech recognition, experiments on a fieldwork corpus of japhug (trans-himalayan family). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178, Dublin, Ireland. Association for Computational Linguistics.

Séverine Guillaume, Guillaume Wisniewski, Benjamin Galliot, Minh-Châu Nguyên, Maxime Fily, Guillaume Jacques, and Alexis Michaud. 2022b. Plugging a neural phoneme recognizer into a simple language model: a workflow for low-resource setting. In *Proc. Interspeech 2022*, pages 4905–4909.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2022. *Glottolog 4.7*. Max Planck Institute for Evolutionary Anthropology.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97.

Sarah Samson Juan, Laurent Besacier, Benjamin Lecouteux, and Mohamed Dyab. 2015. Using resources from a closely-related language to develop asr for a very under-resourced language: A case study for iban. In *Proceedings of INTERSPEECH*, Dresden, Germany.

Sarah Samson Juan, Laurent Besacier, and Solange Rossato. 2014. Semi-supervised G2P bootstrapping and its application to ASR for a very under-resourced language: Iban. In *Proceedings of Workshop for Spoken Language Technology for Under-resourced (SLTU)*.

Shreya Khare, Ashish Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. 2021. Low Resource ASR: The Surprising Effectiveness of High Resource Transliteration. In *Proc. Interspeech 2021*, pages 1529–1533.

Frejus A. A. Laleye, Laurent Besacier, Eugene C. Ezin, and Cina Motamed. 2016. First Automatic Fongbe Continuous Speech Recognition System: Development of Acoustic Models and Language Models. In *Federated Conference on Computer Science and Information Systems*.

Yajie Miao, Florian Metze, and Shourabh Rawat. 2013. Deep maxout networks for low-resource speech recognition. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 398–403. IEEE.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Bhargav Pulugundla, Murali Karthick Baskar, Santosh Kesiraju, Ekaterina Egorova, Martin Karafiát, Lukás Burget, and Jan Cernockỳ. 2018. BUT System for Low Resource Indian Language ASR. In *The Annual Conference of the International Speech Communication Association (Interspeech)*, pages 3182–3186.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging end-to-end ASR for endangered language documentation: An empirical study on yolóxochitl Mixtec. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1134–1145, Online. Association for Computational Linguistics.

Claytone Sikasote and Antonios Anastasopoulos. 2022. BembaSpeech: A speech recognition corpus for the Bemba language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7277–7283, Marseille, France. European Language Resources Association.

Bao Thai, Robert Jimerson, Raymond Ptucha, and Emily Prud'hommeaux. 2020. Fully convolutional asr for less-resourced endangered languages. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 126–130.

Samuel Thomas, Michael L Seltzer, Kenneth Church, and Hynek Hermansky. 2013. Deep neural network features and semi-supervised training for low resource speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6704–6708. IEEE.

Geoffroy Vanderreydt, François REMY, and Kris Demuynck. 2022. Transfer Learning from Multi-Lingual Speech Translation Benefits Low-Resource Speech Recognition. In *Proc. Interspeech 2022*, pages 3053–3057.

Karel Veselỳ, Arnab Ghoshal, Lukás Burget, and Daniel Povey. 2013. Sequence-discriminative training of deep neural networks. In *Interspeech*, pages 2345–2349.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-End Speech Processing Toolkit. In *Proceedings of Interspeech*, pages 2207–2211.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Second to last section.

☑ A2. Did you discuss any potential risks of your work?
Ethics section, last section.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☑ B1. Did you cite the creators of artifacts you used?
If by "artifacts" you mean "datasets", then yes, they are all cited when they are first mentioned.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In the Ethics section we mention that we downloaded some datasets that are publicly available. We also discuss the artifacts that we used that are not publicly available but were shared by indigenous communities with the authors.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Ethics section.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Our own data from indigenous communities was collected under our IRBs. The other data was downloaded from OpenSLR. We explain in the Ethics section that we assume that data was collected ethically but we cannot confirm it ourselves.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Extensively in the data section and ethics sections of our paper.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Section 4, I think.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Vaguely in section 4.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4, 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

We use speech data that was collected and transcribed as part of earlier projects, some by us and some by other groups.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Ethics section

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Ethics

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.