How Important is a Language Model for Low-resource ASR?

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

Nitin Venkateswaran
University of Florida
venkateswaran.n@ufl.edu

Éric Le Ferrand
Boston College
leferran@bc.edu

Emily Prud'hommeaux
Boston College
prudhome@bc.edu

Abstract

N-gram language models (LMs) are the innovation that first made large-vocabulary continuous automatic speech recognition (ASR) viable. With neural end-to-end ASR architectures, however, LMs have become an afterthought. While the effect on accuracy may be negligible for English and Mandarin, jettisoning the LM might not make sense for the world's remaining 6000+ languages. In this paper, we investigate the role of the LM in low-resource ASR. First we ask: does using an n-gram LM in decoding in neural architectures help ASR performance? While it may seem obvious that it should, its absence in most implementations suggests otherwise. Second, we ask: when an n-gram LM is used in ASR, is there a relationship between the size of the LM and ASR accuracy? We have discovered that gut feelings on this question vary considerably, but there is little empirical work to support any particular claim. We explore these questions "in the wild" using a deliberately diverse set of 9 very small ASR corpora. The results show that: (1) decoding with an n-gram LM, regardless of its size, leads to lower word error rates; and (2) increasing the size of the LM appears to yield improvements only when the audio corpus itself is already relatively large. This suggests that collecting additional LM training text may benefit widely-spoken languages which typically have larger audio corpora. In contrast, for endangered languages where data of any kind will always be limited, efforts may be better spent collecting additional transcribed audio.

1 Introduction

Including an n-gram language model (LM) during decoding within an automatic speech recognition (ASR) system is a long-established critical practice (Lucassen and Mercer, 1984; Jelinek, 1976; Bahl et al., 1983, 1989). This has motivated appreciable research exploring different ways to build or augment LMs in order to improve ASR accuracy (Celebi et al., 2012; Bellegarda, 2004; Sagae et al., 2012; Sheikh et al., 2024).

With recent advances in end-to-end neural ASR, the role of the LM has diminished somewhat dramatically (Hannun et al., 2014; Watanabe et al., 2018), with the LM becoming largely optional. Using an LM in decoding is often challenging within many state-of-the-art deep learning ASR toolkits, where it is typically a poorly documented command-line option or a feature that users themselves must implement (Conneau et al., 2020b).

A number of studies have incidentally shown that decoding with an LM within s.o.t.a. neural architectures yields competitive or superior performance (Wang et al., 2022; Liu et al., 2019; Baevski and Mohamed, 2020; Meng et al., 2021; Javed et al., 2022), but the impact of the LM in these papers is discussed briefly, if at all. Additionally, the languages in these studies – typically English – are not at all constrained by data availability. Recent work by Javed et al. (2022) on 40 Indian languages with more limited resources shows a consistent improvement in accuracy when using an LM, but the corpora are still enviably large, with over 17K hours of audio and billions of words.

Two recent studies probe the effect of LM size on ASR performance in more genuinely underresourced settings. Sikasote and Anastasopoulos (2022a) report that a larger LM actually yields mildly worse performance for a 19-hour Bemba corpus. Liu et al. (2023b) explore the effect of LM size in six small corpora, finding a lack of word error rate (WER) improvement when using larger LMs, even in simulated low-resource settings.

The contradictory results in this prior work raise two questions. First, is an n-gram LM still a valuable component in ASR when resources are limited? Second, if using an LM does make a difference, how much additional text data beyond the transcripts of the audio training data should be used to train the LM – or in other words, how large

Findings of the Association for Computational Linguistics ACL 2024, pages 206–213
August 11-16, 2024 ©2024 Association for Computational Linguistics

should the LM be? Training on additional texts could decrease the out-of-vocabulary (OOV) rate and reduce the overall sparsity of the LM, but the impact of the LM *and* its size on ASR accuracy remains strangely unexplored with few studies addressing the issue directly or thoroughly (Sikasote and Anastasopoulos, 2022a; Liu et al., 2023b).

This study aims to provide empirical evidence to answer these research questions. We explore the impact of decoding with LMs of different sizes within two ASR architectures on a diverse set of small corpora. We very *deliberately* choose to work with many corpora of varying sizes, recording quality, and linguistic properties rather than simulating a range of sizes using a single corpus. We conduct further regression analyses to explore for each language whether ASR performance is impacted by LM size and other factors of interest.

Addressing these questions has important implications for creating ASR datasets for under-resourced languages. For many indigenous and endangered languages, additional texts for training larger LMs often come from a domain (e.g., the Bible) that is drastically different from that of the audio (e.g., linguistic fieldwork). This sort of domain mismatch might yield richer LMs that poorly model the target data. An improved understanding of the impact of n-gram LMs on ASR accuracy in such settings can help language communities choose how to allocate their limited resources for building more robust ASR systems.

2 Related work

There has been continuous progress over the past decade in advancing ASR technologies for low-resource languages (Thomas et al., 2013; Cui et al., 2014; Shi et al., 2021). Some of these languages, such as Wolof (Gauthier et al., 2016) and Swahili (Gelas et al., 2012), have large speaker populations; with reasonable time and financial support, more data is obtainable. For others, especially indigenous and endangered languages, it is difficult to collect additional data, but ASR can facilitate the creation of this data (Shi et al., 2021; Prud'hommeaux et al., 2021; Bartelds et al., 2023; Le Ferrand et al., 2023).

Our work goes beyond prior work that examines the impact of n-gram LMs on ASR performance (Sikasote and Anastasopoulos, 2022a; Javed et al., 2022; Liu et al., 2023b). First, we study nine languages from six language families, covering a diverse set of typological properties. Second, we compare two ASR architectures (Section 3.3) for each language. This experimental design allows us to see whether the impact of LM size is consistent across both languages and models.

3 Experiments

It is *not* our goal to find the best performing ASR architecture or parameterizations. We are also *not* comparing WER across languages, a meaningless exercise given the range of languages and corpus size. Instead, our goal is to examine, for each language, the effect of (1) decoding with an LM in an end-to-end system, and (2) increasing the size of the LM, in order to see whether similar *qualitative* trends hold across these nine diverse languages.

3.1 Data sources

We use ASR datasets from nine typologically diverse languages, spanning six language families (see Table 1). Among these languages, Bemba (Sikasote and Anastasopoulos, 2022b), Wolof (Gauthier et al., 2016), Swahili (Gelas et al., 2012), Fongbe (Laleye et al., 2016) (Niger-Congo), Iban (Juan et al., 2014, 2015) (Austronesian) are widely-spoken under-resourced languages (Liu et al., 2022). The datasets for these five languages are publicly available and include additional texts. The dataset for Quechua is taken from the 2022 AmericasNLP Workshop Shared Task[1]. Additional texts are sourced from Agić and Vulić (2019); Ortiz Suárez et al. (2019); Conneau et al. (2020b); Wenzek et al. (2020), and Zevallos and Bel (2023). Hupa (Dene/Athabaskan) is a critically endangered language of North America with audio data derived from fieldwork and additional texts from a published grammar (Goddard, 1904). The corpus for Bininj Kunwo, an Arnhem language spoken in the Northern Territory of Australia, and that for Kréyol Gwadloupéyen, a French-based creole spoken on Guadeloupe (Glaude, 2013), consist of fieldwork recordings with additional texts from the Bible. The last three corpora were shared with us by the linguists who collected the data and are not currently publicly available.

3.2 LM settings

We explore three LM training settings, which we refer to as NO_LM, LM_BASE, and LM_LARGE.

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

Language Name	Audio sources	Audio quality	Additional texts	Audio train	Audio test	LM_BASE size	LM_LARGE size	LM ratio
Bemba	books, radio	variable	religious	19h17m	4h49m	97,148	4,711,467	48.50
Wolof	Wiki, Bible	high	Wiki, Bible	13h27m	3h21m	106,563	708,202	6.65
Swahili	news	variable	news, books	7h17m	1h49m	72,979	29,237,493	400.63
Iban	radio, tv	high	news	6h49m	1h42m	57,755	2,140,207	37.06
Fongbe	daily living	high	news, Bible	5h44m	1h26m	45,567	1,035,713	22.73
Quechua	conversation	variable	religion, gov't	2h59m	0h44m	15,484	2,374,371	153.34
Hupa	fieldwork	variable	grammar	1h16m	19m	7,345	48,731	6.63
Kréyol	fieldwork	variable	Bible	59min	15min	8,857	24,336	2.75
Kunwok	fieldwork	variable	Bible	51min	11min	4,660	281,582	60.43

Table 1: Descriptive statistics. Languages are ordered by total audio size. Numerical counts are those in the *the most recently updated* public repositories. LM size is the number of LM training tokens. Ratio is LM_LARGE:LM_BASE.

NO_LM means that a LM is not included in decoding. Both LM_BASE and LM_LARGE are trigram LMs; the former is trained on the transcripts of the acoustic training data, whereas the latter includes those transcripts plus all the additional text data.

3.3 ASR frameworks

We experiment with two ASR architectures (see also Appendix B). **Kaldi DNN** is a hybrid fully connected DNN implemented with the Kaldi toolkit (Panayotov et al., 2015), following the default sequence training parameters from Karel's recipe[2]. Prior work shows reasonable performance from this DNN architecture in low-resource settings (Morris et al., 2022). Training and decoding within Kaldi requires an LM. For **Wav2Vec2**, we fine-tune from the pre-trained Wav2Vec XLSR-53 multilingual model (Conneau et al., 2020a), built upon the Wav2Vec 2.0 framework (Baevski et al., 2020) (see Table 4, Appendix A for training parameters). The test data is decoded without an LM and, using CTC decoding, with the two trigram LMs.[3]

For evaluation, we use random splits (Liu et al., 2023a), which yield more reliable estimates of performance that the "held-out speaker" approach. We randomly divide the dataset into 80:20 training and test sets, three times. We report the WER score averaged over the three splits.

3.4 Regression analysis

To further validate the effect of LM size on WER, we apply regression modeling. For each language, we first calculate the WER of every LM-decoded test utterance, which serves as the outcome variable in the regression. The variable of interest is the LM size, the number of tokens in the LM training corpus. We include perplexity (PPL) and out-of-vocabulary (OOV) rate as control variables. All fixed-effects have interactions with each other. Given that different LM training settings appear to influence the performance of Kaldi and Wav2Vec2 differently (Section 4; Table 2), we fit separate regression models for each. Recall that with the corpora we have, it is not our goal, nor is it appropriate, to compare WER across languages. Rather, we investigate whether the impact of LM size is consistently observed for every language. To that end, we employ a separate regression model for each language-architecture combination. (Every resulting dataset has a reasonable size, ranging from 1,290 for Kréyol to 15,510 for Bemba given each acoustic model.) This setup also means that the coefficient values for the same factors are not comparable across languages. In cases where speaker (or session, for Swahili) information is provided, we include the speaker/session of the utterance as a random effect, leading to a mixed-effects regression structure; otherwise a linear regression model is applied (Hupa, Quechua, Kréyol, and Kunwok).

4 Results

The WER results are presented in Table 2 (see also Table 5, Appendix C). First, it seems that having an LM, regardless of its size, leads to better performance. All WER results with LM_BASE and LM_LARGE are lower than those derived from NO_LM. Secondly, and counterintuitively for some readers, a larger LM does not increase accuracy for all languages. For Bemba and Swahili, a larger LM yields moderate improvement, but in other cases, regardless of the architecture, the WER for LM_LARGE and LM_BASE are comparable, or the WER for LM_LARGE is slightly worse.

Results from the regression analysis lend support

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

[3] Grid search for the α and β LM parameters yields no appreciable WER improvement (Table 5 in Appendix C).

Language	Model	No_LM	LM_Base	LM_Large
Bemba	Kaldi	-	45.71	42.05
	Wav2Vec2	43.38	38.60	37.33
Wolof	Kaldi	-	31.78	31.93
	Wav2Vec2	21.71	12.64	13.80
Swahili	Kaldi	-	32.44	26.08
	Wav2Vec2	31.74	25.92	24.61
Iban	Kaldi	-	14.54	12.95
	Wav2Vec2	41.24	19.92	19.96
Fongbe	Kaldi	-	55.94	60.28
	Wav2Vec2	16.49	13.68	15.88
Quechua	Kaldi	-	63.65	61.94
	Wav2Vec2	81.35	65.73	66.81
Hupa	Kaldi	-	55.83	56.55
	Wav2Vec2	75.92	50.18	49.29
Kréyol	Kaldi	-	87.75	87.94
	Wav2Vec2	74.34	62.80	62.65
Kunwok	Kaldi	-	70.85	71.03
	Wav2Vec2	76.83	54.84	54.51

Table 2: WER results for all languages, ordered by audio corpus size. No_LM is not applicable to Kaldi.

Language	PPL		OOV (%)	
	LM_Base	LM_Large	LM_Base	LM_Large
Bemba	929.41	2517.17	16.26	9.59
Wolof	93.84	73.53	19.61	19.14
Swahili	366.79	583.78	9.60	2.49
Iban	96.56	101.99	3.07	0.94
Fongbe	22.18	112.34	43.35	43.34
Quechua	2034.23	1314.97	34.20	23.68
Hupa	206.96	212.92	36.43	32.37
Kréyol	158.99	178.43	38.67	34.00
Kunwok	177.35	872.32	47.82	40.43

Table 3: Average perplexity and OOV rates of the transcripts of the acoustic test data.

to these observations. When using Kaldi (Table 6 in Appendix D), there is a significant positive coefficient for LM size for all languages, indicating that a larger LM leads to a higher (worse) WER. These findings largely align with Table 2. Why, then, are the average WERs lower for LM_Large for Bemba and Swahili under Kaldi? The regression results suggest a relationship between LM size and WER *when utterance perplexity and OOV rate are controlled for*. Regression results for Wav2Vec2 resonate with patterns from Table 2 as well: while there is a significant negative coefficient for LM size for Bemba and Wolof, we fail to see the same pattern for other languages.

Why then does having a larger LM not improve model performance? We offer a few conjectures. First, we note in Table 1 that the LM size ratio for some languages is quite small (e.g., Hupa: 6.63; Kréyol: 2.75), while for Bemba, Swahili, Iban and Kunwok, where a larger LM does reduce WER, LM_Large is proportionally much higher than LM_Base. The only exception is Quechua, whose LM size ratio ranks third but the WERs between the two settings are roughly the same.

This brings to our second conjecture: the domain mismatch between the transcribed audio sources and the additional texts may decrease ASR accuracy with LM_Large. For Quechua, Hupa, Kréyol, and Kunwok, the test data comes from everyday speech, whereas the additional texts are technical or religious texts. To explore this idea, we approximate domain similarity by measuring the average perplexity of the transcripts of the acoustic test data under the two LMs (the vocabularies for the two LMs were kept the same). As shown in Table 3, the perplexity scores for LM_Large are higher across languages (with exceptions for Wolof and Quechua) to different extents, suggesting that domain mismatch may play a role in the lack of WER improvement with LM_Large.

Finally, we consider the OOV rates of the acoustic test data. OOV rate will logically decrease with a larger LM, which could in turn yield a lower WER score. This seems true for Bemba, Swahili, Iban and Kunwok. For Wolof and Fongbe, LM_Large does not reduce the OOV rate much. On the other hand, for Quechua, Hupa, and Kréyol, the larger LM does not contribute to improved accuracy despite the lower OOV rate. One noticeable feature of these three languages is that they have relatively little audio training data available, which possibly subsumes the potential effect of LM size.

5 Conclusions and Future Work

Our results over nine languages, corroborated with regression analysis, show that: (1) decoding with an LM yields consistently lower WER; (2) a larger LM infrequently improves ASR accuracy. We propose that in order for a larger LM to be helpful, it needs to: (1) be much larger relative to the size of the transcripts of the audio; (2) result in lower OOV rates for the audio test data; (3) be coupled with sufficient audio training data.

We hope these results can be taken as guidance for ASR dataset creation strategies for low-resource languages. For widely-spoken languages, where additional text data is often readily available, resources might be best directed toward gathering that data. For indigenous and endangered languages, limited resources might be better spent ethically increasing the amount of audio data.

In addition to experimenting with more languages, particularly endangered languages, we see designing informative data selection and data augmentation methods for LM building as a fruitful avenue for future work. This in turn can perhaps mitigate the issues of (audio) data limitations particularly for endangered languages.

6 Limitations

We would like to acknowledge two main limitations of our work: the number of languages and the number of model architectures studied. Given that our experiments target low-resource languages, this is naturally constrained by the lack of ASR dataset availability for languages as such. One possible solution is to create artificial low-resource scenarios from large datasets for languages with abundant training data in order to expand the language diversity covered in this work.

In terms of model architectures, here we explore `Kaldi` and `Wav2Vec2 XLSR-53`, two frameworks that have been shown empirically to work well in settings with limited training data. That said, it would be worthwhile in the future to employ other end-to-end systems (Shi et al., 2021) in order to test their limits in building language technology for low-resource languages.

7 Ethics Statement

We include a total of 10 datasets covering nine languages in our study. Of these datasets, six are publicly available. The data for Hupa, Kréyol, and Kunwok are developed in-house through academic relations with the respective speech communities of these languages; applications of these datasets are carefully considered via personal connections with elders and researchers from the communities.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Alexei Baevski and Abdelrahman Mohamed. 2020. Effectiveness of self-supervised pre-training for ASR. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7694–7698. IEEE.

Alexei Baevski, Henry Zhou, Abdel rahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *ArXiv*, abs/2006.11477.

Lalit R Bahl, Peter F Brown, Peter V de Souza, and Robert L Mercer. 1989. A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(7):1001–1008.

Lalit R Bahl, Frederick Jelinek, and Robert L Mercer. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, PAMI-5(2):179–190.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–729, Toronto, Canada. Association for Computational Linguistics.

Jerome R Bellegarda. 2004. Statistical language model adaptation: review and perspectives. *Speech communication*, 42(1):93–108.

Arda Celebi, Hasim Sak, Erinç Dikici, Murat Saraçlar, Maider Lehr, E Prud'hommeaux, Puyang Xu, Nathan Glenn, Damianos Karakos, Sanjeev Khudanpur, et al. 2012. Semi-supervised discriminative language modeling for Turkish ASR. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5025–5028. IEEE.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdel rahman Mohamed, and Michael Auli. 2020a. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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*.

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.

Herby Glaude. 2013. Corpus Créoloral. oai: crdo. vjf. cnrs. fr: crdo-gcf. *SFL Université Paris*.

Pliny Earle Goddard. 1904. Hupa texts. In *University of California Publications in American Archaeology and Ethnology*, 2, pages 89–368. University Press.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. 2022. Towards building asr systems for the next billion users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10813–10821.

Frederick Jelinek. 1976. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556.

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)*.

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*.

Éric Le Ferrand, Fabiola Henri, Benjamin Lecouteux, and Emmanuel Schang. 2023. Application of speech processes for the documentation of Kréyòl Gwadloupéyen. In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 17–22.

Alexander H Liu, Hung-yi Lee, and Lin-shan Lee. 2019. Adversarial training of end-to-end speech recognition using a criticizing language model. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6176–6180. IEEE.

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3933–3944, Dublin, Ireland. Association for Computational Linguistics.

Zoey Liu, Justin Spence, and Emily Prud'hommeaux. 2023a. Investigating data partitioning strategies for crosslinguistic low-resource ASR evaluation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 123–131, Dubrovnik, Croatia. Association for Computational Linguistics.

Zoey Liu, Justin Spence, and Emily Prud'Hommeaux. 2023b. Studying the impact of language model size for low-resource ASR. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 77–83, Remote. Association for Computational Linguistics.

John Lucassen and Robert Mercer. 1984. An information theoretic approach to the automatic determination of phonemic baseforms. In *ICASSP'84. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 9, pages 304–307. IEEE.

Zhong Meng, Naoyuki Kanda, Yashesh Gaur, Sarangarajan Parthasarathy, Eric Sun, Liang Lu, Xie Chen, Jinyu Li, and Yifan Gong. 2021. Internal language model training for domain-adaptive end-to-end speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7338–7342. IEEE.

John Morris, Justin Chiu, Ramin Zabih, and Alexander Rush. 2022. Unsupervised text deidentification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4777–4788, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7)*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

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.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. 15:491–513.

Kenji Sagae, Maider Lehr, E Prud'hommeaux, Puyang Xu, Nathan Glenn, Damianos Karakos, Sanjeev Khudanpur, Brian Roark, Murat Saraclar, Izhak Shafran, et al. 2012. Hallucinated n-best lists for discriminative language modeling. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5001–5004. IEEE.

Imran Sheikh, Emmanuel Vincent, and Irina Illina. 2024. Training RNN language models on uncertain ASR hypotheses in limited data scenarios. *Computer Speech & Language*, 83:101555.

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 Yoloxóchitl 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. 2022a. Bembaspeech: A Speech Recognition Corpus for the Bemba Language. In *Proceedings of the Language Resources and Evaluation Conference*, pages 7277–7283, Marseille, France. European Language Resources Association.

Claytone Sikasote and Antonios Anastasopoulos. 2022b. 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.

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 *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6704–6708. IEEE.

Yiming Wang, Jinyu Li, Heming Wang, Yao Qian, Chengyi Wang, and Yu Wu. 2022. Wav2vec-switch: Contrastive learning from original-noisy speech pairs for robust speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7097–7101. IEEE.

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.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Rodolfo Zevallos and Nuria Bel. 2023. Hints on the data for language modeling of synthetic languages with transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12508–12522, Toronto, Canada. Association for Computational Linguistics.

A Parameterization for `Wav2Vec2`

The parameterization for `Wav2Vec2` is presented in Table 4.

B Details of Computing Infrastructures

All models are trained on research computing clusters. For `Kaldi`, each model is trained on a single Nvidia A100 GPU with 8GB of memory, and `Wav2Vec` with 16GB of memory.

C Full WER Results

The WER results from `Wav2Vec2` with grid search are presented in Table 5.

D Regression Results

Regression results given different model architectures for our variable of interest, LM size (the number of tokens used to train the language model), and the two control variables, the perplexity (PPL) and OOV are presented in Table 6 (`Kaldi`) and Table 7 (`Wav2Vec2`).

Parameter	**Value**
Number of Epochs	30
Early Stopping Patience	7
Training Batch Size	16
Evaluation Batch Size	8
Warmup Size	1/10 of total stepsize
Gradient Accumulation Size	2
Learning Rate	3e-4 (3e-5 for Iban, Quechua, Hupa)

Table 4: Parameters used to train Wav2Vec XLSR-53.

Language	**Model**	**LM_Base**	**LM_Large**
Bemba	`Wav2Vec2`	38.19	37.02
Wolof	`Wav2Vec2`	12.90	13.36
Swahili	`Wav2Vec2`	25.83	26.40
Iban	`Wav2Vec2`	20.17	19.87
Fongbe	`Wav2Vec2`	14.16	16.04
Quechua	`Wav2Vec2`	68.61	67.82
Hupa	`Wav2Vec2`	50.50	49.18
Kréyol	`Wav2Vec2`	60.67	60.67
Kunwok	`Wav2Vec2`	54.33	54.33

Table 5: WER results from grid search for the dataset(s) of all languages, ordered by the amount of audio data in total. Note that we only performed grid search for when an LM is included for decoding; grid search in this case is not applicable to Kaldi.

Language	**Factor**	**Coef.**	**95% CI**	**Language**	**Factor**	**Coef.**	**95% CI**
Bemba	LM size	1.97	(1.85, 2.08)	Fongbe	LM size	2.44	(2.33, 2.55)
	PPL	0.25	(-0.002, 0.51)		PPL	5.70	(5.10, 6.29)
	OOV	0.46	(0.38, 0.53)		OOV	0.21	(0.18, 0.23)
Wolof	LM size	3.57	(3.51, 3.62)	Quechua	LM size	5.99	(3.66, 8.32)
	PPL	-0.21	(-4.22, 3.46)		PPL	2.88	(-1.85, 7.61)
	OOV	0.23	(0.20, 0.27)		OOV	1.10	(0.26, 1.94)
Swahili	LM size	1.44	(1.40, 1.48)	Hupa (verified)	LM size	2.57	(1.25, 3.89)
	PPL	0.85	(0.73, 0.97)		PPL	-1.18	(-3.75, 1.39)
	OOV	0.14	(0.06, 0.23)		OOV	0.27	(-0.16, 0.70)
Iban	LM size	2.61	(2.28, 2.93)	Kréyol	LM size	4.92	(2.23, 7.61)
	PPL	-4.69	(-5.72, -3.67)		PPL	-0.38	(-5.73, 4.97)
	OOV	-1.60	(-3.49, 0.30)		OOV	0.37	(-3.55, 1.09)
Hupa (coarse)	LM size	9.75	(8.40, 11.09)	Kunwok	LM size	0.02	(-0.15, 1.97)
	PPL	1.70	(-1.16, 4.56)		PPL	-0.67	(-1.03, -3.05)
	OOV	1.92	(1.42, 2.41)		OOV	0.05	(0.01, 0.08)

Table 6: Mixed-effects regression results for each language with WER results derived from `Kaldi`. CI stands for Confidence Interval. A significantly positive coefficient value indicates that the factor leads to a higher WER.

Language	**Factor**	**Coef.**	**95% CI**	**Language**	**Factor**	**Coef.**	**95% CI**
Bemba	LM size	-0.01	(-0.02, -0.003)	Fongbe	LM size	0.01	(-0.01, 0.02)
	PPL	0.01	(-0.01, 0.03)		PPL	0.21	(0.14, 2.82)
	OOV	-0.001	(-0.004, 0.01)		OOV	0.004	(0.0004, 0.01)
Wolof	LM size	-0.01	(-0.01, -0.002)	Quechua	LM size	0.02	(-0.02, 0.06)
	PPL	-0.03	(-0.06, -0.01)		PPL	0.10	(0.02, 0.18)
	OOV	-0.01	(-0.01, -0.002)		OOV	0.02	(0.01, 0.04)
Swahili	LM size	-0.01	(-0.01, -0.003)	Hupa (verified)	LM size	0.04	(-0.05, 0.13)
	PPL	0.04	(0.03, 0.05)		PPL	0.16	(-0.02, 0.34)
	OOV	-0.06	(-0.07, -0.05)		OOV	0.02	(-0.01, 0.05)
Iban	LM size	-0.01	(-0.02, 0.002)	Kréyol	LM size	-0.12	(-0.46, 0.22)
	PPL	-0.03	(-0.07, 0.01)		PPL	-0.20	(-0.88, 0.48)
	OOV	0.005	(-0.07, 0.08)		OOV	-0.03	(-0.12, 0.06)
Hupa (coarse)	LM size	0.02	(-0.07, 0.11)	Kunwok	LM size	-0.01	(-0.06, 0.03)
	PPL	0.09	(-0.10, 0.28)		PPL	0.01	(-0.09, 0.11)
	OOV	-0.02	(-0.05, 0.02)		OOV	0.01	(-0.002, 0.02)

Table 7: Mixed-effects regression results for each language with WER results derived from `Wav2Vec2`. CI stands for Confidence Interval. A significantly positive coefficient value indicates that the factor leads to a higher WER.