Studying the impact of language model size for low-resource ASR

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

Justin Spence
University of California, Davis
jspence@ucdavis.edu

Emily Prud'hommeaux
Boston College
prudhome@bc.edu

Abstract

This paper investigates the impact of language model (LM) size on low-resource ASR, using data from five widely-spoken low-resource languages and one endangered Native American language. Our findings demonstrate that having larger LMs does not necessarily result in lower WER; this is most evident for the endangered language, where larger LMs actually led to significantly worse performance than that observed in the widely-spoken low-resource languages. We conjecture that one of the potential driving forces behind this discrepancy is the domain mismatch between the transcripts of the audio data and the supplementary texts used to train the LM. We discuss the implications of our results in the context of creating ASR corpora for low-resource languages.

1 Motivation

The language model (LM) has long been noted as an important component for the decoding process of automatic speech recognition (ASR) (Bahl et al., 1989; Chelba et al., 2012; Sak et al., 2012; Arisoy et al., 2015; Park et al., 2019). The texts used for training LMs usually consist of transcripts of the audio training data along with additional texts from other sources, such as Web text, literature, or transcripts of previously recorded audio (see Section 3). When building an ASR corpus, researchers typically make efforts to gather large amounts of these supplementary texts in order to improve LM coverage and reduce out-of-vocabulary (OOV) rates.

For high-resource languages, considerable amounts of supplementary text are relatively easy to find and process. Acquiring additional data for low-resource languages can be more challenging. For some widely spoken low-resource languages, such as Swahili, which has millions of speakers and an established writing system, additional textual data can be acquired from websites and documents that have been digitized (Liu et al., 2022a). Thus, when building ASR corpora for widely spoken low-resource languages, researchers usually collect additional text data (Laleye et al., 2016; Juan et al., 2014; Gauthier et al., 2016).

The situation for endangered languages is much more complex. Collecting additional text for endangered languages often involves digitizing archival materials held by tribal authorities or archiving organizations. Processing these materials requires manual oversight and linguistic expertise, particularly when there are multiple conflicting orthographic traditions.

While obtaining additional texts for widely spoken low-resource languages may also be labor-intensive, it is more straightforward to ensure that the additional texts will belong to domains similar to that of the audio transcripts. In contrast, for endangered languages, particularly those of North America, the additional texts are often from restricted domains such as religious documents, grammar textbooks, or fieldwork materials from decades ago (Liu et al., 2022b). The subject matter and linguistic characteristics of these texts can be quite different from those of the recordings usually used for ASR acoustic training, which are typically recent recordings of elders telling stories or having conversations. These differences can lead to substantial *domain mismatch*.

These limitations on acquiring additional LM training text for low-resource languages raises the question: how important is additional text for improving low-resource ASR? Or put differently: is it necessary to take the time to acquire and process additional text to build a larger LM?

This study takes up these questions with data from five widely-spoken low-resource languages and one endangered language. We investigate to what extent and in what context LMs impact ASR performance. Given the different challenges for widely-spoken vs. truly low-resource languages, we expect the role of the LM to be different. It is

Language	Data sources		Audio			Additional written texts	
	Audio	Additional texts	Train	Test	N of words in audio training data	N of words	Proportion
Fongbe	daily living phrases	news; Bible	5h44m	1h26m	45,544	990,146	21.74
Wolof	read speech (Wikipedia/ Bible)	exactly same source as audio	15h11m	3h47m	121,220	601,639	4.96
Swahili	news	news; books	8h47m	2h11m	84,287	29,237,493	346.88
Iban	radio/television station	news	6h49m	1h42m	57,608	2,082,452	36.15
Bemba	literature; radio	religious magazines	19h40m	4h55m	106,657	4,614,319	43.26
Hupa (verified)	the elder's stories	grammar books	1h16m	19m	7,369	41,386	5.62
Hupa (coarse)	same as verified data	same as verified data	6h6m	1h31m	32,640		1.27

Table 1: Descriptive statistics for audio data and additional written texts used to train language models for each language in the experiments. Note that the numerical counts here were derived directly from the public repositories and may be different from those originally reported in the papers. **Proportion** refers to the relative ratio between N of words in additional texts and N of words in the transcripts of the audio training data.

possible that adding additional text to build larger LMs for an endangered language might not actually lower WER as much as one might expect, if at all.

2 Related Work

A number of studies have focused on providing ASR data sets and models for languages that have many speakers but limited existing ASR training resources (see Section 3), such as Fongbe (Laleye et al., 2016) and Iban (Juan et al., 2014). For languages that lack both resources and large numbers of speakers, there have also been efforts utilizing ASR technologies to support their documentation (Adams et al., 2018; Jimerson and Prud'hommeaux, 2018; Gupta and Boulianne, 2020b,a). For instance, Shi et al. (2021) built end-to-end models for the Oto-Manguean language, Yoloxóchitl Mixtec. Zahrer et al. (2020) studied phoneme recognition for the Muyu language from the Trans–New Guinea language family.

Another line of relevant work attends to investigating effective evaluation methods for low-resource ASR, in a way to strive for more generalizable model performance. Comparing three model architectures, including both neural and non-neural alternatives, across five languages, Morris et al. (2021) pointed out that no model architecture is a clear "winner"; rather the results for model rankings depend on the language. Liu et al. (forthcoming) explored how different data split methods influence WER scores for under-resourced scenarios using five languages as the test cases. Their findings demonstrated that the commonly-applied "held-out speakers" evaluation scheme for ASR falls short when there exists high speaker variability in the data set; in other words, the performance of ASR systems for low-resource languages, at least in the context that they investigated, is not *speaker-independent*.

3 Meet the data

We used ASR corpora for five widely-spoken low-resource languages, which are publicly available: Fongbe (Laleye et al., 2016), Wolof (Gauthier et al., 2016), Swahili (Gauthier et al., 2016), Iban (Juan et al., 2014), and Bemba (Sikasote and Anastasopoulos, 2022). We also included one data set developed for Hupa, a critically endangered language indigenous to North America. Details of the sources for the transcripts of the audio data and the additional texts gathered to train the LM are presented in Table 1. We present additional

information for the Hupa language below.

The audio recordings for Hupa came from linguistic fieldwork which started in 2005 and is still ongoing. The recordings were produced by a single female elder speaker from the speech community, a scenario that is unfortunately common for critically endangered languages. Transcription of these recordings included several stages of manual correction, and ambiguous or unclear audio was confirmed with the elder before being considered as complete. Hence, some transcripts have been checked more thoroughly than others. Depending solely on the differences of transcription quality, the recordings and their transcripts were divided into two sets, which we will refer to as "verified" vs. "coarse" data respectively.

Looking across the data described here, it seems likely that there is less domain overlap between the audio data and the additional LM-training text for Hupa than for the other languages. For Fongbe, Swahili, Iban and Bemba, both the audio and the additional text are related to news or local radios; for Wolof, the audio data and the additional texts are from the exact same source. For Hupa, the grammar book data used for LM training are more formal in style and has little overlap in content with the contemporary fieldwork recordings.

4 Experiments

We probe whether the assumption that having a larger LM (i.e., an LM built on more words of training text) leads to lower WER will hold for low-resource languages. A recent study of ASR for Bemba (Sikasote and Anastasopoulos, 2022) compared the impact of two LMs on WER, one of which was 29 times larger than the other. The results showed that, surprisingly, the larger LM led to slightly worse performance than the smaller one. We sought to address whether the data of other languages demonstrates the same pattern. Specifically, we carried out different experimental settings (basic and simulated settings). In each setting we explored different configurations of the training texts for the LM in order to investigate the role of LM size on ASR performance.

4.1 Basic settings

For the data set(s) of each language, in the basic setting, we first created two different LMs, `LM_base` and `LM_large`: the former was built using *only* the transcripts of the audio training data, while training for the latter also included *all* the additional LM training text.

In addition to comparing ASR performance in a given language with and without supplementary LM training text, which can vary considerably in size, we also investigated the impact of the LM training corpus size relative to the size of the corpus of transcriptions. Note that for the coarse data set of Hupa, the size of the additional texts is only 1.27 times of that of the transcripts of the audio training data (Table 1). This is proportionally smaller by comparison to other languages as well as to the verified data set of Hupa. Therefore, for each of the other data sets, except for the coarse data of Hupa, we kept the audio training and test data the same, then randomly sampled (3 times) sentences from the additional texts such that the ratio between the size of these texts and that of the transcripts of the audio training data also approximates 1.27. These sampled sentences along with the transcripts of the audio training data were then combined to build what we refer to as a proportionally-sized LM, `LM_prop`.

4.2 Simulated settings

Given that the amount of audio training data is different for each language, it is possible that even when the LM size is the same proportionally, the WER results are dependent on the amount of audio. With that in mind, we also explored simulated settings to ensure that more fair conclusions could be drawn, especially when comparing widely-spoken vs. truly low-resource languages.

Here we focused on Fongbe, Iban, and Swahili (which have the smallest audio dataset sizes compared to Wolof and Bemba). Again, since Hupa has the least amount of either audio data (verified data set) or additional LM training text by proportion (coarse data set), our simulated settings involve data subsampling from each of the other three languages above in order to construct augmented data sets whose *audio training data size* is similar to that of the two data sets for Hupa. Hence, for each language, we created a verified and a coarse setting. (It would be ideal to create these two settings such that the audio data quality difference between the two mimic that between the verified and the coarse data sets of Hupa; however, it is not exactly clear how this can be achieved in a principled way, making it beyond the scope of this work.)

Take Iban as an example. Recall that each record-

ing for every language has been manually segmented into utterances. For the verified setting, we first randomly sampled (3 times) a number of utterances such that the total duration of these utterances was similar to the total duration of audio of the verified data set for Hupa. We then created two different LMs, `LM_base` and `LM_prop` for the sampled audio data in the same way as we did for the basic settings described above. For the coarse setting, we carried out the same procedure except that the total duration of the sample audios was similar to that for the coarse data set of Hupa.

4.3 LMs and acoustic models

All LMs are trigram LMs trained with Witten-Bell discounting using the SRILM (Stolcke, 2002) toolkit. To build acoustic models, we used the open-source Kaldi toolkit (Povey et al., 2011). Specifically, we adopted a recipe of a fully connected deep neural network (DNN) from Kaldi with the default sequence training parameters; this model architecture has six hidden layers with 1024 hidden units in each. This architecture has been shown to outperform other statistical alternatives such as subspace Gaussian mixture models, as well as neural models such as the time delay neural networks (Morris et al., 2021; Morris, 2021).

For the data set of every language other than those for Swahili and Hupa, we conducted acoustic feature transformations for each individual speaker. For the data of Swahili, which lacks clear speaker identity information, and the two data sets of Hupa, which only contains recordings from one speaker, we carried out acoustic feature transformation for each recording date or recording session separately. Model training was carried out with state-level minimum Bayes risk criterion and a per-utterance Stochastic Gradient Descent weight update. For decoding, we used the Kaldi finite state transducer-based decoder.

4.4 Evaluation scheme

It is common to perform ASR evaluating using "held-out speaker(s)", namely holding out the data of one set of speakers as the test set and leaving the remaining data for training, without conducting cross-validation (i.e., using the data of different sets of speakers as the test set) (Gauthier et al., 2016; Zeyer et al., 2019). Nevertheless, Liu et al. (forthcoming) found this evaluation scheme to be problematic in that the performance of the acoustic models is *dependent on which speaker(s) were included in the test set*. Alternatively, they proposed using random splits, presenting strong evidence that the average WER across all held-out speakers is comparable not only to the average WER derived from multiple random splits of the full acoustic data, but also to the WER of *just one random split*.

Here we also adopted random splits for ASR evaluation. Specifically, the audio data of each language was randomly split into training/test sets (3 times) such that their respective total utterance duration approximates a ratio of 4:1.[1]

5 Results

We present the results from the basic settings in Table 2. Note that we are not trying to compare WER scores across languages, as they are evidently not comparable. Instead, our focus is to compare WER scores *derived from different LM sizes for each individual language*, then examine whether the effect of LM size is in the same direction across the data sets of the languages studied here. (While we attend to WER here, we also calculated character error rate (CER) as an additional evaluation metric; the patterns of CER largely follow those of WER). Table 2 suggests that having an LM built on a larger text data set does not always lead to lower WER. For widely-spoken low-resource languages, having larger LMs (both `LM_prop` and `LM_large`) resulted in lower WER for Iban, Bemba and Swahili. On the other hand, the WER scores became mildly worse for Fongbe and Wolof when the LM size was larger. Particularly for Hupa, the truly low-resource language studied here, larger LMs had a negative impact; this is most evident in the case of the coarse data set, where `LM_large` actually increased WER score by 37.73% compared to `LM_base`.

To further examine whether LM size potentially has a different effect on Hupa compared to other languages, and that the effects are not necessarily caused by other languages having more audio data, let us turn to the results from the simulated settings. As demonstrated in Table 3, in both the verified and coarse settings for most languages except for Fongbe, larger LM size resulted in lower WER

[1] Note that the ASR corpora of most languages here provide a lexicon file (required by Kaldi), possibly extracted from external dictionaries, for the decoding process of the acoustic models. Since we are interested in the size of LMs, we tried to control for additional factors as much as possible. Therefore every time a new model was to be trained, we generated a lexicon file directly from the corresponding LM.

Language	LM_base	LM_prop		LM_large	
	WER (CER)	WER (CER)	WER (CER) reduction	WER (CER)	WER (CER) reduction
Fongbe	59.81 (0.36)	60.44 (0.37)	-1.05 (-2.78)	61.7 (0.36)	-3.16 (0)
Wolof	28.75 (0.15)	29.11 (0.14)	-1.25 (6.67)	29.41 (0.14)	-2.3 (6.67)
Swahili	30.24 (0.13)	28.59 (0.13)	5.46 (0)	25.35 (0.11)	16.17 (15.38)
Iban	14.8 (0.06)	14.8 (0.06)	0 (0)	13.53 (0.05)	8.58 (16.67)
Bemba	46.09 (0.12)	44.38 (0.12)	3.71 (0)	42.82 (0.11)	7.09 (8.33)
Hupa (verified)	54.06 (0.31)	56.42 (0.29)	-4.37 (6.45)	54.83 (0.32)	-1.42 (-3.23)
Hupa (coarse)	43.68 (0.22)	-	-	60.16 (0.31)	-37.73 (-40.91)

Table 2: Evaluation results from basic settings; *reduction* (%) refers to WER (CER) reduction compared to the WER (CER) (%) when using `LM_base`.

Language	Setting	LM_base	LM_prop	
		WER/CER	WER/CER	reduction
Fongbe	verified	66.29 (0.42)	67.65 (0.46)	-2.05 (-9.52)
Fongbe	coarse	-	-	-
Swahili	verified	54.86 (0.28)	50.84 (0.26)	7.33 (7.14)
Swahili	coarse	30.42 (0.14)	28.26 (0.13)	7.1 (7.14)
Iban	verified	30.07 (0.13)	26.93 (0.12)	10.44 (7.69)
Iban	coarse	15.38 (0.06)	14.41 (0.06)	6.31 (0)

Table 3: Evaluation results from simulated settings; note that course setting does not apply to Fongbe given its audio data size; *verified* and *coarse* refer to the simulated settings following the setup of Hupa, and do not refer to the quality of the data; *reduction* (%) refers to WER (CER) reduction compared to the WER (CER) (%) when using `LM_base`.

instead, indicating better model performance.

These findings have two implications. First, they suggest that the influence of LM size on ASR performance varies between widely-spoken and truly low-resource languages, and that larger LMs are more likely to have negative effects for the latter, at least in the settings that we investigated. One possible explanation for this discrepancy is that described in Section 3, namely that there often exists a more substantial domain mismatch between the transcripts of the audio data and the additional LM texts for endangered languages, an important factor that was not mitigated by simply having more training texts for LMs. Second, comparing the results from Table 2 and Table 3, for widely-spoken low-resource languages in particular, it seems that the impact of LM size could also interact with the size of the audio data, in the sense that when the amount of audio data is small, having a larger LM tends to have more positive influence on ASR performance.

6 Discussion and Conclusion

With data from five widely-spoken low-resource languages and one endangered language of North America, we studied the impact of LM size on ASR performance. Our results demonstrate that, perhaps surprisingly, having larger LMs does not always result in lower WER. This observation is the most pronounced for the truly low-resource (endangered) language in contrast to the widely-spoken low-resource languages. In addition, our findings suggest that the effect of LM size is potentially modulated by the amount of audio data available; larger LMs more consistently lead to better model performance when the amount of audio data is relatively small.

The aforementioned observations indicate there would be value in collecting additional texts to build ASR corpora for widely-spoken low-resource languages. However, they raise questions about the utility of such endeavors for endangered languages, when the domain of these texts might be very different from that of the audio transcripts. For future work, it would be worthwhile to study a wider set of typologically diverse languages with varying sizes for the LMs, in order to assess how the languages' phonological and morphological properties might potentially affect ASR performance. Additionally, one should carry out experiments comparing the impact of LMs whose training texts are explicitly from different domains, which would help further confirm the influence of domain mismatch. Relevant findings could in turn inform the creation of ASR corpora for low-resource languages broadly.

Acknowledgements

We are grateful for the continued support from the Hupa indigenous community. We would like to especially thank Mrs. Verdena Parker for her generous and valuable input for the documentation work of Hupa throughout the years. In addition, we thank the anonymous reviewers for their helpful feedback. This material is based upon work supported by the National Science Foundation under Grant #2127309 to the Computing Research Association for the CIFellows Program, and Grant #1761562. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the Computing Research Association.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluation phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ebru Arisoy, Abhinav Sethy, Bhuvana Ramabhadran, and Stanley Chen. 2015. Bidirectional recurrent neural network language models for automatic speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5421–5425. IEEE.

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.

Ciprian Chelba, Dan Bikel, Maria Shugrina, Patrick Nguyen, and Shankar Kumar. 2012. Large scale language modeling in automatic speech recognition. *arXiv preprint arXiv:1210.8440*.

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).

Vishwa Gupta and Gilles Boulianne. 2020a. Automatic transcription challenges for Inuktitut, a low-resource polysynthetic language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2521–2527, Marseille, France. European Language Resources Association.

Vishwa Gupta and Gilles Boulianne. 2020b. Speech transcription challenges for resource constrained indigenous language Cree. 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 362–367, Marseille, France. European Language Resources association.

Robbie Jimerson and Emily Prud'hommeaux. 2018. ASR for documenting acutely under-resourced indigenous languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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*.

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022a. 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. forthcoming. 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: Main Volume*.

Zoey Liu, Justin Spence, and Emily Tucker Prud'hommeaux. 2022b. Enhancing documentation of Hupa with automatic speech recognition. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 187–192, Dublin, Ireland. Association for Computational Linguistics.

Ethan Morris. 2021. Automatic Speech Recognition for Low-Resource and Morphologically Complex Languages. Master's thesis, Rochester Institute of Technology.

Ethan Morris, Robert Jimerson, and Emily Prud'hommeaux. 2021. One size does not fit all in resource-constrained ASR. In *The Annual Conference of the International Speech Communication Association (Interspeech)*, pages 4354–4358.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*, pages 2613–2617.

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.

Haşim Sak, Murat Saraçlar, and Tunga Gungor. 2012. Morpholexical and discriminative language models for Turkish automatic speech recognition. *IEEE transactions on audio, speech, and language processing*, 20(8):2341–2351.

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. 2022. 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.

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Alexander Zahrer, Andrej Zgank, and Barbara Schuppler. 2020. Towards building an automatic transcription system for language documentation: Experiences from Muyu. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2893–2900, Marseille, France. European Language Resources Association.

Albert Zeyer, Parnia Bahar, Kazuki Irie, Ralf Schlüter, and Hermann Ney. 2019. A comparison of Transformer and LSTM encoder decoder models for ASR. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 8–15. IEEE.