Enenlhet as a case-study to investigate ASR model generalizability for language documentation

Éric Le Ferrand
Boston College
leferran@bc.edu

Raina Heaton
University of Oklahoma
rainaheaton@ou.edu

Emily Prud'hommeaux
Boston College
prudhome@bc.edu

Abstract

Although both linguists and language community members recognize the potential utility of automatic speech recognition (ASR) for documentation, one of the obstacles to using these technologies is the scarcity of data necessary to train effective systems. Recent advances in ASR, particularly the ability to fine-tune large multilingual acoustic models to small amounts of data from a new language, have demonstrated the potential of ASR for transcription. However, many proof-of-concept demonstrations of ASR in low-resource settings rely on a single data collection project, which may yield models that are biased toward that particular data scenario, whether in content, recording quality, transcription conventions, or speaker population. In this paper, we investigate the performance of two state-of-the art ASR architectures for fine-tuning acoustic models to small speech datasets with the goal of transcribing recordings of Enenlhet, an endangered Indigenous language spoken in South America. Our results suggest that while ASR offers utility for generating first-pass transcriptions of speech collected in the course of linguistic fieldwork, individual vocabulary diversity and data quality have an outsized impact on ASR accuracy.

1 Introduction

The fields of descriptive and documentary linguistics concentrate on collecting and analyzing language samples, particularly from understudied, Indigenous, and endangered languages. Typically, documentary linguists – who can be researchers or language community members – make audio recording of unscripted or prompted speech, followed by transcription, glossing, translation, and analysis. Transcription, however, often becomes bottleneck when dealing with large speech corpora, rendering only a fraction of the available speech data available for analysis or for language instruction (Himmelmann, 1998).

Automatic speech recognition (ASR) has emerged as a potential solution by providing first-pass transcripts that can be manually corrected (Mitra et al., 2016; Bird, 2021; Jimerson and Prud'hommeaux, 2018). The preferred approach for building an ASR system with scarce resources is to fine-tune a large multilingual model to whatever small amount of transcribed audio data is available for the target language. Many demonstrations of the efficacy of this approach, however, rely on corpora with relatively few speakers or with recordings made under the same condition (e.g., all read speech or broadcast news) (Jimerson et al., 2023). One problem is that models trained on a single uniform speech corpus may overfit to that corpus acoustically or lexically. It is not clear how such models will generalize to new data – whether that data is archival recordings or recent data from a different speaker population or data collected with different prompts.

In this paper, we address this question using a corpus of speech recordings for the language Enenlhet (ISO 639-3 code `tmf`; not to be confused with the related language Enlhet, ISO 639-3 code `enl`). Enenlhet is spoken by fewer than 2,000 people living in what is now Paraguay. While thousands of the world's languages, like Enenlhet, are endangered and have minimal written documentation, many of these languages lack three important characteristics that make Enenlhet an ideal language for exploring the utility of ASR for documentation of diverse speech data. First, the amount of available transcribed speech data – 5 hours – is relatively substantial for an endangered and primarily oral language. Second, the large quantity of untranscribed audio – over 100 hours – is highly unusual for any endangered language, offering the potential for unsupervised training and for experimentation with integrating ASR into the documentation pipeline. Third, the Enenlhet speakers who provided their voices have stated enthusiasm for generating new

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 132–137
June 21, 2024 ©2024 Association for Computational Linguistics

documentation for their language in collaboration with outsiders.

Using two ASR architectures that support fine-tuning acoustic models to the task of ASR for small speech datasets, we explore to what extent an existing corpus can be used to train models that generalize well to new data. The dataset we use, while part of a single data collection effort, was collected over multiple years under varying conditions from a large number of speakers. We simulate introducing a new recording by holding out each speaker in turn, training on the remaining speakers, and testing on the held-out speaker. We find that the high degree of lexical diversity across speakers, as well as differences in audio and transcription quality, contribute to variability in word error rate (WER), findings we quantify with a regression analysis.

2 Related work

ASR has long been proposed as a solution to the "transcription bottleneck" challenge of language documentation, but there has been relatively little effort dedicated to practically using ASR for this purpose. The focus of much of the early work was phone-level transcription (DiCanio et al., 2013; Johnson et al., 2018; Zahrer et al., 2020). Other applications have involved keyword spotting (Le Ferrand et al., 2022) or the development of front-end tools for building ASR systems (Foley et al., 2018). Only recently has ASR been actually used in an active documentation pipeline (Prud'hommeaux et al., 2021; Gupta and Boulianne, 2020; Shi et al., 2021; Rodríguez and Cox, 2023). There is some prior work investigating data partitioning strategies (Liu et al., 2023), which we indirectly address in our work when we use a held-out-speaker approach to simulate testing a trained model on new data. Le Ferrand et al. (2023) applied a trained model for an under-resourced language to new data, yielding surprisingly weak results and indicating that models fine-tuned on small amounts of data may not generalize well to new data. We also note previous work on the impact of specific dataset characteristics, including OOV rate and audio quality which we explore here, on word error rate (Jimerson et al., 2023). This last paper includes data from the AmericasNLP 2022 ASR shared task (Ebrahimi et al., 2022). While the AmericasNLP datasets were extremely small (typically less than one hour), they contained fieldwork recordings with characteristics similar to those included in our study.

3 Data

The language of the corpus used in this study is known by a number of names (Cabanatit, Enenxet, Toba-Enenlhet), but following the preferences of the community, we will refer to it here as Enenlhet (ISO 639-3 code tmf). Enenlhet is a polysynthetic language spoken by fewer than 2,000 people living in the Paraguayan Chaco region. Migration and displacement have led to dramatic language loss; the current Ethnologue status of Enenlhet is 6b (Threatened). Enenlhet has remarkably little available documentation. Aside from a few short word lists compiled in the 1920s and 1960, there are no dictionaries, and there is only one available grammar (Unruh et al., 2003). A phone set of approximately 4 vowels and 15 consonants can be inferred from the dataset described below.

The data used here is part of a recent multi-year data collection effort, which has so far yielded over 120 hours of recordings with more than 40 individuals. The data was collected with university IRB approval and is archived with the Archive of Indigenous Languages of Latin America at the University of Texas. Ethical practices require a consultation with the language community before using language material for research purposes (Pirinen et al., 2024). Thus, while the data we use is publicly available[1], the co-author who collected the data gained express permission from the community for research purposes. She notes that the Enenlhet speakers who participated in the data collection were eager for their speech recordings to be used to support documentation and revitalization efforts.

Approximately 10 hours of the recordings from 16 speakers have been transcribed with utterance-level timestamps. The total quantity of speech data available in these recordings – after stripping out silence, segments in another language, and speech produced by the interviewer – is approximately 5 hours. Table 1 shows information for each speaker.

4 Methods

4.1 Experiments

We trained ASR models using two frameworks that support fine-tuning from a multilingual model: Whisper (Radford et al., 2022) and wav2vec (Baevski et al., 2020; Conneau et al., 2021). In the case of Whisper, we used the Whisper medium

[1] https://ailla.utexas.org/islandora/object/ailla%3A266554

	SSA	CA	ER	IF	PA	OM	HM	FF	TF	MRR	MM	AR	LM	BT	MR	LF
duration	80:16	65:00	5:40	20:32	18:51	26:18	10:45	39:00	5:20	12:21	4:53	7:00	4:21	2:42	3:00	2:48
tokens	6663	5828	434	1495	1571	2139	756	3192	362	868	509	814	349	209	253	201
types	2181	1769	228	513	847	965	437	1448	229	399	261	214	164	134	144	117

Table 1: Duration (MM:SS), token count, and type count for each of the 16 speakers in our dataset.

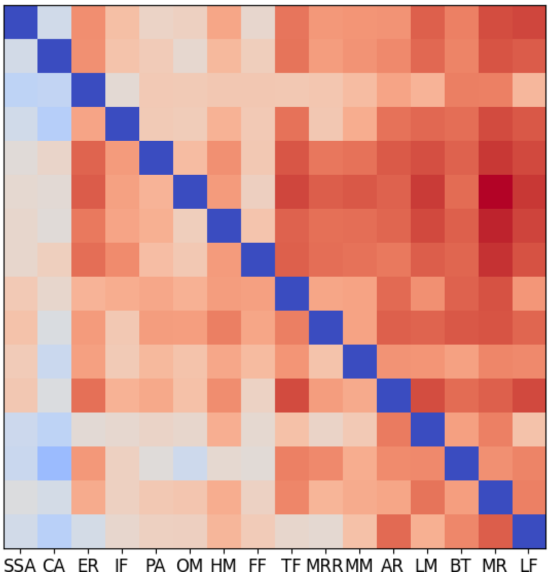

Figure 1: OOV heatmap: Blue indicates low OOV rate while red indicates high OOV rate, with the hue indicating more extreme values in each respective direction.

model, adhering to the hyperparameters specified in the main tutorial[2]. For wav2vec, we employed xlsr-53, following the hyperparameters of the main tutorial[3]. Regarding wav2vec decoding, we decoded with a language model trained on the transcripts of the relevant training data. Notably, we opted for default values for decoding parameters α and β given their minimal impact with small LMs.

Recall that the recordings used were collected over several years, in different recording conditions from different individuals. Our goal is not to create a robust ASR system, but rather to assess whether a model trained on existing data will generalize to a new speech recording or corpus. Initially, we train a baseline model by randomly partitioning the entire dataset into training and testing sets. Subsequently, we use a "leave speaker out" cross validation approach to simulate the testing of new data on an existing trained ASR model. For each speaker within our dataset, we train an ASR model using all recordings except for the those of the target speaker, whose data is reserved for testing purposes.

4.2 Analysis

The ASR experiments are evaluated using the traditional Word Error Rate (WER) metric. We then aim to identify factors that could impact system performance. First, we focus on the lexicon, examining two factors: the Out-of-Vocabulary (OOV) rate, which represents the proportion of tokens in the test set that did not appear in the training set (see Figure 1). The blue cells, corresponding to the longest recordings SSA and CA, have a substantially larger vocabulary that overlaps with the rest of the collection. We perform the same analysis with the types. We then calculate the duration of both the training and testing sets. Following this, we assess the audio quality in the test sets based on two measures: loudness and sharpness. Loudness is a measure designed to mimic sound perception in humans, while sharpness relates to the subjective perception of high-frequency content in a sound. Both sharpness and loudness are determined using the Zwicker method (Zwicker, 1960) with the Mosquito toolkit[4]. Finally, we evaluate the transcription quality by conducting a CTC-based alignment of the transcription and utilizing the resulting CTC posterior probabilities as a measure of transcription reliability. Our intuition is that low CTC probabilities indicate that the alignment algorithm had difficulty determining the alignment, perhaps because of noisy recordings or inconsistent transcriptions. To perform the alignment we used an ASR model trained in English (wav2vec2_ASR_base_960h).

5 Results

Results are presented in Figure 2. Each bar corresponds to a test conducted on a different speaker's data. The baselines are indicated by the dotted lines. First, we see across all scenarios, wav2vec performs systematically worse than Whisper. Second, we observe in all experiments, the baseline does not exhibit consistent inferior or superior performance in either architecture. We note that potential biases during experiments conducted on random splits do not significantly impact overall performance in one

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

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

[4] https://github.com/Eomys/MoSQITo/tree/master

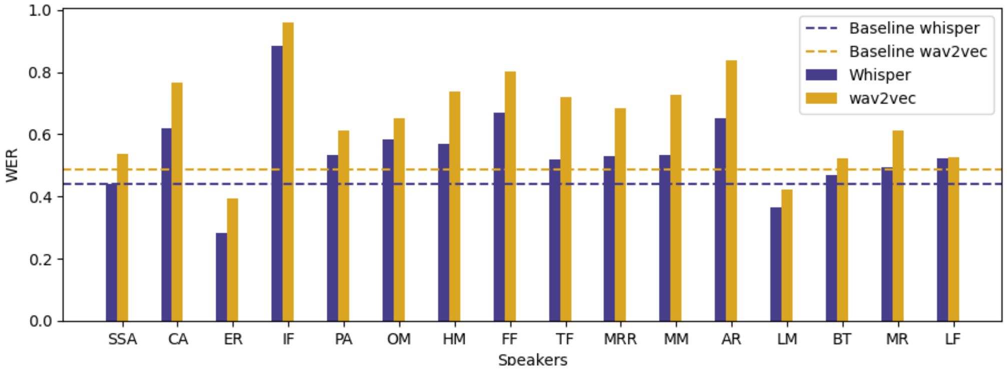

Figure 2: WER across all speakers. Baselines are derived from a random split across all speakers.

	OOV tokens	OOV types	train duration	test duration	loudness	sharpness	alignment score
Coeff.	0.79	0.36	-0.001	0.001	0.001	0.134	-1.588
95% CI	(0.07, 1.1)	(0.07,1.4)	(0.07,184.)	(0.07,12.3)	(0.07, 0.67)	(0.07,2.04)	(0.07,18.9)

Table 2: regression results with WER results derived from Whisper. CI stands for Confidence Interval.

	OOV tokens	OOV types	train duration	test duration	loudness	sharpness	alignment score
Coeff.	1.171	0.452	-0.001	0.001	0.000	0.301	-2.547
95% CI	(0.08, 1.2)	(0.08,1.6)	(0.08,210.)	(0.08,14.)	(0.08, 0.76)	(0.08,2.25)	(0.07,20.5)

Table 3: regression results with WER results derived from wav2vec. CI stands for Confidence Interval.

direction or the other. Secondly, performance is dependent on factors inherent to the test speaker's data. The regression analysis enabled us to ascertain the impact of these factors on WER.

The results of the regression analysis can be found in Table 2 for Whisper and Table 3 for wav2vec. A significantly positive coefficient value indicates that the factor leads to a higher (worse) WER while a significantly negative coefficient indicates that the factor leads to a lower (better) WER. Across both architectures, three factors do not have a significant influence on WER: duration of the training data and test data, and loudness. One of the most relevant factors is the OOV tokens and to a lesser extent, the OOV types. These factors happen to be much more salient for wav2vec than for Whisper which can perhaps explained by the use of a language model during decoding in wav2vec.

Two metrics were used evaluate the quality of the audio in the test sets. Loudness was found to have minimal impact, but sharpness appears to negatively impact WER. Additional experiments on larger datasets are necessary to validate the efficacy of this measure in assessing audio quality.

The posterior probabilities obtained from the CTC alignment exhibited a strong negative impact on WER, suggesting that a higher confidence score in the alignment corresponds to a lower WER. However, the confidence interval is relatively high, raising doubts about the reliability of this measure to evaluate the quality of the transcription. Examining specific examples, we verified the data quality for IF, where a very high WER was observed. It was discovered that the transcription for this speaker did not align with the audio; instead, it appeared to be a translation of the audio into a related language or dialect. However, the CTC alignment did not substantially differ from other speakers where the transcription matched well and the WER is much lower. This measure appears instead to be relevant for evaluating audio quality when there is not a significant mismatch with the transcription.

6 Conclusions

This paper explores how contemporary speech recognition architectures perform in a language documentation setting, focusing on the Enenlhet language as a case study. In order to simulate testing of new data using a model trained on a previous data collection corpus, we conducted training and testing of ASR models using a leave-one-out evaluation approach, where the models were trained on all Enenlhet speakers except one and tested on the one left out. Additionally, we performed a regression analysis to determine the factors that may influence WER.

The experimental results initially revealed that the leave-one-out evaluation approach neither outperforms or underperforms a random split approach for our specific case. Subsequently, we found that Out-of-Vocabulary (OOV) rates are the most significant factor in explaining the WER for a given test set. Lastly, both the sharpness measure and the CTC posterior probabilities show promise in assessing the quality of the speech signal, which could potentially correlate with the word error rates. Further analysis is necessary to confirm this correlation. These results suggest that in low-resource settings, ASR models may not always generalize well to new data, which could hamper the utility of ASR for language documentation.

References

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.

Steven Bird. 2021. Sparse transcription. *Computational Linguistics*, 46(4):713–744.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Proceedings of Interspeech*, pages 2426–2430.

Christian DiCanio, Hosung Nam, Douglas H Whalen, H Timothy Bunnell, Jonathan D Amith, and Rey Castillo García. 2013. Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *The Journal of the Acoustical Society of America*, 134(3):2235–2246.

Abteen Ebrahimi, Manuel Mager, Adam Wiemerslage, Pavel Denisov, Arturo Oncevay, Danni Liu, Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues, et al. 2022. Findings of the second americasnlp competition on speech-to-text translation. In *NeurIPS 2022 Competition Track*, pages 217–232. PMLR.

Ben Foley, Joshua T Arnold, Rolando Coto-Solano, Gautier Durantin, T Mark Ellison, Daan van Esch, Scott Heath, Frantisek Kratochvil, Zara Maxwell-Smith, David Nash, et al. 2018. Building speech recognition systems for language documentation: The coedl endangered language pipeline and inference system (elpis). In *SLTU*, pages 205–209.

Vishwa Gupta and Gilles Boulianne. 2020. 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.

Nikolaus P. Himmelmann. 1998. *Documentary and descriptive linguistics*, volume 36. de Gruyter.

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)*.

Robert Jimerson, Zoey Liu, and Emily Prud'hommeaux. 2023. An (unhelpful) guide to selecting the best asr architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1008–1016.

Lisa M Johnson, Marianna Di Paolo, and Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing prosodylab-aligner with tongan data.

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.

Éric Le Ferrand, Steven Bird, and Laurent Besacier. 2022. Learning from failure: Data capture in an Australian aboriginal community. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4988–4998.

Zoey Liu, Justin Spence, and Emily Prud'hommeaux. 2023. 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.

Vikramjit Mitra, Andreas Kathol, Jonathan D Amith, and Rey Castillo García. 2016. Automatic speech transcription for low-resource languages-the case of yoloxóchitl mixtec (mexico). In *INTERSPEECH*, pages 3076–3080.

Flammie Pirinen, Linda Wiechetek, Trond Trosterud, Sjur Moshagen, and Børre Gaup. 2024. Computel partnerships in practice: Giellalt. In *Proceedings of 7th the Workshop on Computational Methods for Endangered Languages*.

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

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*.

Lorena Martín Rodríguez and Christopher Cox. 2023. Speech-to-text recognition for multilingual spoken data in language documentation. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 117–123.

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.

Ernesto Unruh, Hannes Kalisch, and Manolo Romero. 2003. *Enenlhet apaivoma: nentengiai'a nengiangveiakmoho neliatekamaha enenlhet apaivoma, guía para el aprendizaje del idioma materno toba*. Nengvaanemkeskama Nempayvaam Enlhet.

Alexander Zahrer, Andrej Žgank, and Barbara Schuppler. 2020. Towards building an automatic transcription system for language documentation: Experiences from muyu. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2893–2900.

Eberhard Zwicker. 1960. Ein verfahren zur beredinung der lautstärke. *Acta Acustica united with Acustica*, 10(4):304–308.