Customizing ASR for Language Documentation and Resource Prioritization

Alexandra Fort and Shobhana Chelliah
Indiana University, Bloomington, IN USA

Abstract

Research in language documentation has the potential to benefit from integration of ASR models, especially through the assisted transcription of recordings with audio. Recent advancements in ASR for low-resource languages demonstrate the ability to adapt general, multilingual models for unseen languages with limited fine-tuning data, supporting the creation of custom ASR models. However, resources are still required to collect and prepare the fine-tuning data, necessitating exploration of optimization of resource allocation within the process of data collection and preparation. This paper outlines important considerations for the collection and preparation of data for customizing an ASR model for use in language documentation projects. With the development of a Lamkang ASR model as an example, prioritization of tasks within a language documentation project is outlined by analyzing the relative impact of time spent on transcription correction versus time spent on manual alignment on ASR model performance. Results from this research suggest prioritization of transcription correction over manual-alignment of data and suggest fine-tuning multilingual ASR systems produces superior results to zero-shot ASR models, despite recent advancements in the technology.

1 Introduction

In language documentation projects, the most valuable resource is commonly the time that the language experts, both linguists and speakers, spend refining the language data in a project. In the documentation process, a corpus of the language is developed to provide a foundation for research on the language and to guide the development of linguistic resources (Himmelmann et al., 2006). Audio recordings of speech acts, such as a conversation, interview, or narrative, are included in the corpus to provide acoustic examples of the language. There recordings are frequently accompanied by transcriptions that depict their contents. The act of transcribing is considered the slowest part of the process of documenting a speech act, producing a phenomenon known as transcription bottleneck, which refers to the idea that the speed of transcription slows the rate at which a documentation project can progress, both due to the time it takes to transcribe and the amount of skilled transcribers available (Bird, 2020).

The idea of a transcription bottleneck is further complicated in language documentation projects where the orthography of the language is being established simultaneously. The process of establishing an orthography is time-intensive, but very important for the language community's ability to interact with the project (Seifart et al., 2006). When orthography development and language documentation are happening concurrently, the initial transcription of a speech act is subject to change, requiring additional editing as decisions about the orthography progress. In such scenarios, the project may decide to either transcribe recordings phonetically, such as through the use of IPA, or through the use of the developing orthography.

There are various pros and cons to each approach. If using a phonetic representation, the transcriptions can be converted to whichever orthography is established after an orthography is established. However, this requires the transcriber to be trained to use a phonetic alphabet. Additionally, the transcribed recordings are likely to be inaccessible to most of the community until being converted to a representation that is being used in the community (Dobrin and Schwartz, 2021). Interpretability by the community is important both for avoiding extractive research practices and creating space for community language experts to contribute their invaluable knowledge to the project. This input during the transcription process is the most significant benefit of using the developing orthography instead of IPA. Further, use of the developing orthography supports familiarization and continued discus-

sion within the community about the orthography, which is required to further formalize the orthography. The primary limitation of this approach lies in the necessitation of updates, as transcripts using a developing orthography require updates. Moreover, if the developing orthography moves from a phonetically-condensed representation (such as using one character to represent multiple phonemes) to a more phonetically-specific representation, the ability to perform automatic conversion of the orthography is limited.

This specific study concerns a project that has chosen the method of using a developing orthography for transcription, detailing observations from the process of trying to develop the best-possible ASR model with the existing data. While low-resource is used to cover an expanse of situations, this particular project includes use of less than 3 hours of language data. The research questions addressed in this paper are:

1. Does time invested in alignment of transcripts or updating the orthography result have more impact on ASR model performance?

2. How does the relationship of data quantity and data quality impact training results in low-resource settings?

3. How do various models designed for use in low-resource settings compare to one another, both in performance and usability?

2 ASR and Language Documentation

ASR-aided transcription is often proposed as a method with the potential to address the transcription bottleneck. Examples of near-perfect performance of transcription tasks for high-resource languages motivate researchers to investigate how these tools could be adapted and applied in lower-resource settings. Previously, traditional ASR systems for low-resource languages made use of phoneme and lexicon dictionaries to produce language-specific systems, but recent research has proven the utility of end-to-end (E2E) systems for the task. Particularly, multilingually-trained E2E systems have been successfully fine-tuned on a small subset of labeled data in many low-resource settings (Olev and Alumae, 2022; Shi et al., 2021; Sadeque et al., 2022; Coto-Solano et al., 2022), both in scenarios where the initial E2E system includes the language in their training data and in scenarios where fine-tuning is the system's first exposure to the language. This approach requires less language-specific preparation of phoneme and lexicon dictionaries, lowering the barrier of preparation required to utilize ASR systems for unseen languages.

Limited research exists on the utility of the generated transcripts or the threshold of performance required by the ASR model to produce a useful transcript. Specifically, research is needed to determine at which point a transcript has too many errors, making error-correction a more difficult task than generating the transcript manually. This is a complicated topic of research as many factors are likely to impact the perceived difficulty of error correction, including transcriber preference, setting of the initial recording, and standardization of the operational orthography. For example, Prud'hommeaux et al. (2021) worked with the Seneca community and found that ASR-assisted transcription is able to produce transcripts in a shorter time period and with fewer errors, but that, regardless, many speakers prefer to do unassisted transcription. There is not a particular character error rate (CER) or word error rate (WER) that has been established as sufficient for aiding transcription, just a general assumption that the lower the rate is, the easier and faster the transcript will be to correct. However, it is also important to note that error-correction tasks come with their own challenges, as the distribution and types of error impact their detection (Søby et al., 2023; Point and Baruch, 2023).

Rendering an audio recording into a transcript necessitates decisions about what should be transcribed from the input. These decisions are of on-going interest for both high-resource and low-resource settings. For example, in high-resource settings, ASR systems have been shown to create accessibility issues for users with speech disfluencies, such as stuttering, by not being able to properly identify what users are saying (Mujtaba et al., 2024). Liao et al. (2023) suggest post-processing of ASR transcripts to reduce the influence of the transcription of disfluencies and speech errors on the readability of the final transcript. However, in language documentation scenarios where transcribers are simultaneously learning the components of speech that are relevant to the linguistic inventory of the language, reduction of such phenomena is complicated. Further, Ko and Burch (2025) emphasize the relevance and importance of maintaining speaker variation in transcription, as variation in transcription can change for both differ-

ent speakers and the same speaker on different days. They note various transcriptions can be produced that are valid representations of the audio, with different considerations for what is being annotated and stylistic decisions.

Results from languages without an established orthography demonstrate that ASR systems are able to produce a phonetically-consistent transcript for an unseen language (Chizzoni and Vietti, 2024). However, depending on the phonetic consistency of a language's orthography, this default ability to produce phonetically-consistent transcripts may not contribute to improved results for a particular language, as not all phonetically apparent components are relevant to speakers. Bird (2020) builds on this issue, asserting that careful transcription of all components of a transcript may be misaligned with a community's priorities and that sparse transcription could be more beneficial for some language documentation projects.

While fine-tuning E2E systems is the popular approach in most of the existing research, being able to use an ASR system without additional fine-tuning would increase access to the technology for many language documentation projects. While using off-the-shelf ASR systems for transcription when the target language is not in the training data has historically produced suboptimal results (Lin et al., 2025; Zahrer et al., 2020; Zheng et al., 2022), the recent Omnilingual model releases have shown improvement for zero-shot ASR (Omnilingual ASR Team et al., 2025), wherein the target language is missing from the training data and no fine-tuning is performed. Specifically, the zero-shot model, `omniASR_LLM_7B_ZS`, is able to use a multilingually trained ASR model plus an LLM to generate a reasonable ASR model for the language with just 1-10 recordings (of up to 30 seconds) provided for context.

However, Omnilingual ASR Team et al. (2025) do show that fine-tuning `Wav2Vec2` models with even just 10 hours of language-specific data outperforms usage of the zero-shot model. Other projects have also seen promising results with less than 10 hours of data for fine-tuning an E2E, such as the use of 4 hours for Cook Island Maori (Coto-Solano et al., 2022) and the use of less than 99 minutes of single-speaker data for Limbu, Dotyal, Duoxo, Nahsta, Mwotlap, and Vatlong (Boulianne, 2022). This suggests that given even a very limited amount of transcribed data, fine-tuning an existing multilingual E2E model is likely to provide improved performance over the zero-shot model. Nonetheless, integrating results from both approaches in this study allows for further discussion of strengths and weaknesses of zero-shot ASR versus fine-tuning multilingual E2E models.

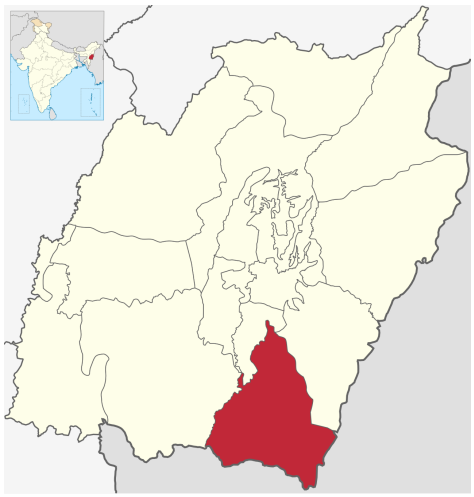

Figure 1: Highlight of the Chandel district within the Manipuri state of India, the primary area where Lamkang is spoken (Commons, 2021)

3 Lamkang

3.1 Language Overview

Lamkang (iso: lmk) is a Tibeto-Burman language in the South Central (formerly known as Kuki-Chin) branch. It is spoken in Manipur, India by about 10,000 people, based on the results from India's 1999 census (Eberhard et al., 2024). The language is primarily spoken in villages in the Chandel District, though recent conflict in the Manipur region has resulted in the relocation of speakers to nearby cities. The language is mentioned in the 1927 Linguistic Survey of India (Grierson, 1927), two grammatical sketches have been written for Lamkang (Thounaojam and Chelliah, 2007), as well as a handful of linguistic articles about reduplication, conjugation, and spatial terminology (Chelliah et al., 2020, 2019; Chelliah and Utt, 2017).

Two NSF grants (#1160640 and #0755471) from 2008-2016 were used to develop a corpus for the language. The corpus includes transcriptions of monologues, discussions, interviews, and elicitation narratives from 6 speakers (Sumshot Khular, Rex Rengpu Khullar, Swamy Tholung Ksen, Daniel Tholung, Shekarnong Sankhil and Kumar Sankhil). The recording environment is partially controlled, but background noise and crosstalk are apparent. Transcriptions are accompanied by the

Updated	mthungbi ava thung thang va, mtii thang ngi, talu daat to boorkaang ne
Original	mbih ava' thung thang vah, mtii thang ngi, talu daat a boor kaang ne
Translation	And then there, beneath there, what was that, a basket, right?

Table 1: A side-by-side of a sentence in the updated orthography and the original orthography. Punctuation added for clarity in translation.

original recording audio, with a limited number being time-aligned in ELAN.

Lamkang uses a Latin-based orthography, but is still in the process of formalizing the orthography. Chelliah et al. (2023) details orthographical variation in Lamkang based on speaker writings collected during 12 years of language documentation work. Comparing this variation to orthographical decisions made in related languages and considering language-specific circumstances, suggestions are made for a series of orthographical decisions, such as segmentation and vowel length. Based on this analysis, Lamkang orthography uses 39 graphemes to represent vowels and consonants, as shown in the table found in the appendix 3. While Lamkang is a tonal language, currently the orthography does not mark tone. Analysis of the tonal system of Lamkang is ongoing in the documentation progress, so discussion of specifics related to Lamkang tone is left to future work.

3.2 Lamkang ASR Data

The current Lamkang project contains 38 transcribed files. The audio includes monologues, discussions, interviews, and elicitation narratives (such as Pear Stories). The files are primarily in Lamkang, though some of the interviews include questions posed in English. Additionally, instances of Lamkang-English code-switching, commonly singular words borrowed from English into Lamkang grammar structures, are found in some of the recordings. The files are between 12 seconds long and 20 minutes 26 seconds long with an average recording length of 5 minutes and 3 seconds. Of the 38 files, 19 have been aligned manually while another 19 are unaligned. The manually-aligned files have been aligned at phrasal levels in chunks varying in length from 1-12 seconds.

The initial transcriptions of the files were done by linguists in conjunction with community members. However, the orthography has evolved since the initial transcriptions and the initial transcriptions require updates based on recently established orthographic conventions, primarily related to word boundaries, representations of vowel length, and inclusion of phonemes that have been reduplicated or inserted at word boundaries due to phonological rules. 8 of the 38 files have been updated to reflect the latest orthography. Compared to the original, if considering the latest orthography the gold-standard, there is a character error rate (CER) of .106 between the 8 updated transcripts and the 8 original transcripts. Variations in the initial transcriptions make automatic methods of updating the transcriptions to the existing standard unreliable, meaning that original transcriptions require review by a language expert or linguist. This leaves two primary tasks for data preparation of the files: transcription review and alignment.

4 Methodology

Considering the two data preparation tasks for the Lamkang data, transcription review (updating the orthography) and alignment, significant time is required to prepare the data for ASR. However, as time is a costly resource, this research methodology is meant to explore the most efficient allocation of time in this context, specifically whether updating transcriptions or manually aligning data provides a larger boost to the performance of a Lamkang ASR model, as well as how this relates to the amount of training data. Specifically, should researchers focus on correcting the original transcripts to the current standard, depending on tools for automatic alignment, or would it be better to correct and manually align a subset of the original transcripts? This question is intrinsically tied to the performance of auto-alignment methods, as improvements in automatic alignment methods should reduce the dependence of projects on manual alignment. Additionally, it contributes to broader discussions about data quantity versus quality.

With the goal of investigating the most productive path forward, 6 Lamkang ASR models are fine-tuned or tested, as shown in table 2. The first 5 models are fine-tuned versions of `Wav2Vec2` (Baevski et al., 2020) while the last model in the table represents the results of in-context learning (ICL) using the latest released Omnilingual ASR model, `omniASR_LLM_7B_ZS` (Omnilingual ASR

Team et al., 2025). Hyperparameters for fine-tuning can be found in Appendix B. The fine-tuned versions of `Wav2Vec2` are the primary focus for this paper due to the consensus about improved performance reported in section 2, but the `omniASR_LLM_7B_ZS` Omnilingual model was released during this research and results are presented for comparison of performance and relative ease of use.

For naming shorthand, Man and Auto are used to refer to manual versus automatic alignment while C and O refer to use of corrected versus original transcripts. Com is shortened to indicate combined and Omni to indicate the Omnilingual model. Of the `Wav2Vec2` fine-tuned models, the ManC model is trained on 7 recordings that have been manually reviewed by both a linguist and a native speaker and manually-aligned, representing prioritization of thoroughly cleaned and reviewed data at the expense of quantity. The ManO model includes the same 7 manually-aligned recordings, but retains the original transcriptions, showing the impact of updating the orthography. The AutoC model uses the 7 corrected transcripts, but replaces the manual alignment with automatic alignment, allowing for assessment of manual alignment versus automatic. Lastly, the AutoO model uses 37 of the original transcripts and automatic alignment with the goal of evaluating whether data quantity is more important than quality for ASR performance.

The ComC model uses the 7 corrected recordings, but uses both the manually-aligned and automatically-aligned transcripts to test the impact of a data augmentation method. This differs slightly from a simplistic duplication of the fine-tuning data by introducing the model to different prosodic chunks of the audio. For the OmniC model, 10 examples can be provided to the base model. For this set of 10, the 7 longest clips from each of the 7 corrected, manually-aligned recordings were selected, plus the 3 next longest clips when looking at all of the clips in aggregate. Documentation suggests use of clips of less than 30 seconds for the context examples, but when a set of clips that were about 20 seconds long on average were used for training, the transcript produced repetition loops. Providing a smaller subset of 5 recordings with the same average duration of 20 seconds long produced the same error, suggesting that using longer durations of audio for ICL of the Omnilingual model does not improve the results and may trigger a common LLM failure mode, such as repetition loops. The average duration of the clips used in the reported model is 6.55 seconds, as this set of clips did not produce the repetition loop error.

The total length of the set of audio clips used for fine-tuning or ICL is reported in the third column of 2. Note that small discrepancies in the total time of the audio clips of the ManC, ManO, and AutoC models is a result of differing segmentations and pauses between words in those different segmentations. Code for data preparation, fine-tuning, and testing models is found at `https://github.com/aconeil/lamkang_asr`.

5 Data Preparation

For the AutoC and AutoO models, the first step of data preparation was to automatically-align the files using the accompanying text transcripts. An initial review of the text files, both corrected and original, included removing any speaker diarization marks, general punctuation cleanup, and references to indistinct noise or uncertainty[1]. While many of these aspects are handled later on in the ASR fine-tuning pipeline, cleanup was required to compare the updated and original transcriptions and support correct segmentation for automatic alignment. Manual correction was used in place of automatic cleanup of the text files to protect against accidental deletion of transcript-relevant phenomena. Automatic alignment was done by resampling the audio to 16000Hz and using the MMS forced alignment model (MMS_FA) (Pratap et al., 2024) with PyTorch. This produced word-level alignment for the transcripts, with white-space in the transcripts delimiting words. The alignments were exported to ELAN files using the speach Python library.

For all `Wav2Vec2` fine-tuned models, the ELAN files were split by annotation segments to produce `wav` files for training the model. An accompanying `csv` file was produced for the split annotations that maps the annotations to each `wav` file. The automatic word-level alignments are too short for fine-tuning ASR models, so the words are grouped into 7-gram long segments with remainder words in a transcript being appended to the final recording clip of the file. This methodology means that the manually-aligned segmentations are more likely to follow human-intuition and account for suprasegmental features, such as prosody and tone, while

[1] Sequences were marked variably in transcripts and referred to different phenomena (i.e. mumbling, fast speech, trailing speech)

the automatically-aligned segmentations are split without regard to these features. Selection of the clips for the OmniC model follows the description provided in 4 section, with clips coming from the processing and segmentation of the corrected, manually-aligned files described above.

6 Results

The evaluation of these models uses character error rate (CER) and word error rate (WER). Since CER represents the normalized edit distance between two strings by calculating the insertions, deletions, or substitutions needed to turn one string into another and the next step of using an automatic transcription tool is to have a proficient speaker or linguist review and edit the output, CER can be used to approximate the edits required by the reviewer. WER is also provided, as updates to tokenization in Lamkang are part of the ongoing process of standardizing the orthography and the cognitive effort of identifying word breaks is likely to impact the speed of transcript correction.

In order to compare the performance of the models to one another, the test set is the same for all of the models. Additionally, since the goal in a documentation project would be to conform to the most updated orthography, the test set comes from one of the corrected transcripts. The test set consists of one recording with updated orthography, coming from a recording that is 1 minute and 47 seconds long. Accordingly, all versions of the transcription of this file and audio are removed from fine-tuning sets, including the transcription of the audio in the older version of the orthography.

The relative performance of these models suggests that updating the orthography is more beneficial to model performance than manual alignment, at least when the updates to the orthography are around a 10% CER, as is the case with this dataset. Using original orthography and manual alignment resulted in a model with a WER and CER that were respectively 11.8% and 3.7% worse, while using updated orthography and automatic alignment produced a model with a WER and CER that were 1.7% and 3.5% worse. Though the updated orthography has about a 10% CER when compared to the original orthography, the CER between the AutoC and ManO models is similar. However, there is a more than 10% improvement in WER in the AutoC model, demonstrating the utility of using the updated orthography to facilitate correct tokenization.

The best-performing model, the ComC model, achieves the lowest CER and WER, showing that automatic alignment can be leveraged for model improvement when used as an augmentation method. Compared to just using the manually-aligned and corrected recordings, the addition of automatically-aligned and corrected recordings gave a WER improvement of 8.7% and a CER improvement of 1.9%. Though the significant decrease between performance in the ManC and AutoC models shoes that using automatic alignment in place of manual alignment decreases performance, it is beneficial when used in conjunction with the manually-aligned data.

In this research, the automatic alignment model consisting of all of the original transcripts (AutoO) is unable to properly fine-tune, regardless of adjustments to hyperparameters. In the best configuration of hyperparameters, very early stopping (at step 500) could result in a AutoO model with a CER of around .80, suggesting an underlying issue with the data itself in that the addition of data is actually hurting the model's performance almost immediately. Through error analysis and manual review of the automatically-aligned ELAN files, a significant issue that arose with automatic alignment was attribution of non-speech sounds to words in the transcript. This was especially problematic for recordings that began with or included a longer period of background noise. Trimming the beginning of such audio files and performing noise reduction would likely result in improved alignment, but removing these elements from the recordings would also diminish the ability of the model to appropriately handle such occurrences in recordings. Generally, more review of the transcripts would be required to make them useful for automatic alignment, though it is not clear that these steps would fully address this issue. Further, it is not clear manual cleaning of the recordings would be significantly faster than manual alignment.

Lastly, the OmniC model has a reasonably low CER at .268, but a high WER at .815. As the model required minimal resources and efforts for ICL, the CER is impressive at .268. The high WER but low CER indicates that tokenization is a bottleneck for the `omniASR_LLM_7B_ZS` model, though phoneme identification performs relatively well. Though this project is in a low-resource setting, providing 13 minutes of fine-tuning data to `Wav2Vec2` is still able to produce superior results to the `omniASR_LLM_7B_ZS` model, results that mirror

Name	Description	Duration	WER	CER
ManC	Manual-align, corrected	00:13:18	.575	.205
ManO	Manual-align, original	00:13:16	.693	.242
AutoC	Auto-aligned, corrected	00:14:36	.592	.240
AutoO	Auto-aligned, original	02:58:13	1	1
ComC	Both, corrected	00:27:54	.488	.186
OmniC	Manual-align, corrected	00:01:05	.815	.268

Table 2: Description of models trained and total duration of wav files used for training (HH:MM:SS), followed by their character error rate (CER) and word error rate (WER)

the findings of the model release paper (Omnilingual ASR Team et al., 2025).

7 Discussion

Regarding the discussion of data quality versus data quantity, the model results showcase how determining which to prioritize is situation dependent. The ManO model outperforming the ManC model provides an example of quantity eclipsing quality. However, as alignment issues arise with the AutoO model, we see that the quality of the data can be detrimental to other parts of the ASR pipeline. Automatic alignment presents some utility for data augmentation, especially when used with the corrected transcripts, as seen by the ComC model, but is not able to produce adequate results with the current state of the original data.

If the original transcription data is very clean, such as list elicitation in a well-controlled environment, it is possible automatic alignment would be more beneficial to the project. For continuing research on tone identification, establishing a pipeline that utilizes list elicitation in a highly controlled environment and using automatic alignment could allow for efficient processing and integration of tonal data and mitigate the issues seen with non-speech sounds in automatic alignment of the existing transcripts. As for the more naturalistic settings (interviews, story-telling, and discussion) found in these transcripts, automatic alignment will be unlikely to produce data that is ready for fine-tuning an ASR model due to idiosyncrasies of trancsribers and small deviations between transcript and audio that naturally occur when transcribing longer audio files with more background noise. It is possible that automatic alignment at the word level could serve as a starting point for someone manually aligning the transcripts, with the caveat that adjusting the annotation boundaries and merging words may or may not be quicker for a transcriber than manually selecting prosodic chunks.

Considering the process of fine-tuning the models, the Omnilingual model was much easier to use. It does still require programming, but the bar to entry is much lower in terms of data preparation and data availability. This approach would likely only be recommended in language documentation situations in which there is no transcribed data in the language, but a linguist and/or speaker are able to transcribe 10 short recordings to get the process started. At this point, there is the possibility that the transcripts produced by the Omnilingual could be used as a starting point for further transcription via error correction, though, as mentioned in section 2, additional research is needed to determine at which point an ASR-produced transcript has a CER that is sufficiently low enough for the process of error correction to be faster than starting from scratch. Further investigation of the utility and application of the Omnilingual model at the nascent phase of a language documentation project would provide further insight as to the model's utility, but the model does not apply well for the needs of this project.

8 Conclusion

Based on the initial results of this research, this project would see a greater benefit from updating the orthography than manually aligning the transcripts. While both are important to improving the results of an ASR system, greater gains in WER and a comparable CER is seen when including updates to the orthography compared to using the original manually-aligned data. However, each project must consider the resources available for review. For example, manual-alignment of data requires less knowledge of the language and specialized attention than updating the orthography, so this approach may be a better option if those available for review have less time and language-

specific expertise.

When considering data quality and data quantity, we see the importance of the two fluctuate depending on the degree of automation involved in the pipeline. If depending on additional tools, such as automatic alignment, data quality has a heavier influence, but when considering quality in terms of orthographic updates, quantity was more influential. Noting the limitations of automatic alignment based on data quality, we see the potential to improve model performance when the method is used with cleaned, high-quality data, such as the corrected transcripts.

Lastly, in considering the ease of use and performance of the various models, the results support preceding research in finding that significant gains in performance are found when fine-tuning E2E ASR models with limited labeled data. Though the zero-shot model requires less effort for data preparation and a dramatically reduced quantity of data, the current results of the model are easily surpassed with about 13 minutes of annotated data. The omnilingual model is more likely to benefit the very beginning phases of a documentation project and aid in the process of building enough labeled language data to fine-tune a model. While it presents a impressive step forward in zero-shot ASR, more improvements, especially in tokenization, are required for it to surpass the performance of a fine-tuned, multilingually-trained E2E ASR model.

Limitations

This research presents findings that are specific to the development of an ASR model for the Lamkang language documentation project, so research results are influenced by the specific language and project context. Though these results inform research on low-resource languages, specifics of other languages and projects are necessary to gain a fuller understanding of challenges and strengths of current ASR models for low-resource languages. The data used in the project is very limited, as it comes from an actual language documentation project that is in progress. The research guides the language documentation project on the most efficient way to create more annotated data, which can then be used to increase the sample size and significance of results. However, the current amount of audio available does limit the statistical significance of the results.

Acknowledgments

This research would not be possible without the dedication, work, and insight into the Lamkang language provided by Setpu-One Silsi.

This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Steven Bird. 2020. Sparse transcription. *Computational Linguistics*, 46(4):713–744.

Gilles Boulianne. 2022. Phoneme transcription of endangered languages: an evaluation of recent asr architectures in the single speaker scenario. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2301–2308.

Shobhana Chelliah, Evaline Blair, Melissa Robinson, Rex Khullar, and Sumshot Khular. 2020. Reduplication in lamkang: Form, function, feeling. *Expressive morphology in the languages of South Asia*, pages 167–186.

Shobhana Chelliah, Rachel Garton, Sumshot Khular, and Rex Khullar. 2023. Orthography development for languages of the south central branch of tibeto-burman: Lessons from lamkang. *Himalayan Linguistics*, 22(1).

Shobhana Chelliah, David Peterson, Tyler Utt, Evaline Blair, and Sumshot Khular. 2019. Lamkang verb conjugation. *Himalayan Linguistics*, 18(1).

Shobhana Lakshmi Chelliah and Tyler P Utt. 2017. The syntax and semantics of spatial reference in lamkang verbs. *Himalayan Linguistics*, 16(1).

Ilaria Chizzoni and Alessandro Vietti. 2024. Towards an asr system for documenting endangered languages: A preliminary study on sardinian. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 214–220.

Wikimedia Commons. 2021. Chandel in manipur (india).

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, Liam Koka'ua, Syed Tanveer, and Isaac Feldman. 2022. Development of automatic speech recognition for the documentation of cook islands māori. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3872–3882.

Lise Dobrin and Saul Schwartz. 2021. The social lives of linguistic legacy materials. *Language Documentation and Description*, 21.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. *Ethnologue: Languages of the World: Lamkang*, 27th edition. SIL International, Dallas, Texas.

George Abraham Grierson. 1927. *Linguistic survey of India*. Office of the superintendent of government printing, India.

Nikolaus P Himmelmann et al. 2006. Language documentation: What is it and what is it good for. *Essentials of language documentation*, 178(1).

Edwin Ko and Jem Burch. 2025. Transcription as an iterative and interpretive practice: Documenting connected speech in apsáalooke (crow). *Language Documentation and Description*, 25(1).

Junwei Liao, Sefik Eskimez, Liyang Lu, Yu Shi, Ming Gong, Linjun Shou, Hong Qu, and Michael Zeng. 2023. Improving readability for automatic speech recognition transcription. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(5):1–23.

Kaiying Kevin Lin, Hsiyu Chen, and Haopeng Zhang. 2025. Formosanbench: Benchmarking low-resource austronesian languages in the era of large language models. *arXiv preprint arXiv:2506.21563*.

Dena Mujtaba, Nihar Mahapatra, Megan Arney, J Yaruss, Hope Gerlach-Houck, Caryn Herring, and Jia Bin. 2024. Lost in transcription: Identifying and quantifying the accuracy biases of automatic speech recognition systems against disfluent speech. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4795–4809.

Aivo Olev and Tanel Alumae. 2022. Estonian speech recognition and transcription editing service. *Baltic Journal of Modern Computing*, 10(3):409–421.

Omnilingual ASR Team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. 2025. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages.

Sébastien Point and Yehuda Baruch. 2023. (re) thinking transcription strategies: current challenges and future research directions. *Scandinavian Journal of Management*, 39(2):101272.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language documentation and conservation*, 15.

Zarif Al Sadeque et al. 2022. *Automatic speech recognition for documenting endangered first nations languages*. Ph.D. thesis, University of Saskatchewan.

Frank Seifart et al. 2006. Orthography development. *Essentials of language documentation*, pages 275–299.

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.

Katrine Falcon Søby, Byurakn Ishkhanyan, and Line Burholt Kristensen. 2023. Not all grammar errors are equally noticed: error detection of naturally occurring errors and implications for eye-tracking models of everyday texts. *Frontiers in Psychology*, 14:1124227.

Harimohon Thounaojam and Shobhana L Chelliah. 2007. The lamkang language: Grammatical sketch, texts and lexicon. *Linguistics of the Tibeto-Burman Area*, 30(1):1–212.

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.

Weiyi Zheng, Alex Xiao, Gil Keren, Duc Le, Frank Zhang, Christian Fuegen, Ozlem Kalinli, Yatharth Saraf, and Abdelrahman Mohamed. 2022. Scaling ASR Improves Zero and Few Shot Learning. In *Interspeech 2022*, pages 5135–5139.

A Lamkang Orthography

The following table displays graphemes in Lamkang.

B Hyperparameters

The following hyperparameters were used during training:

Grapheme	Phoneme	Example	Gloss
a	[a], occasionally [ə]	**a**rh**a**ng k**a**l m**a**	*Don't climb up*
aa	[a:]	prkh**aa**	*almond*
ai	[aj]	ph**ai**vang	*ant*
aai	[a:j]	ps**aai**	*elephant*
ao or au	[a:w]	phk**ao**; **au**va	*reptiles*; *that one*
b	[b]	**b**aak rek	*bats*
ch	[tʃ], ocassionally [ts]	**ch**en	*to run*
d	[d]	**d**ii	*water*
e	[e]	ch**e**t lam da	*they went*
ee	[e:]	mkh**ee**l thung bi ngu	*when they asked*
ei	[ej]	n**ei**	*I*
h	[h]	**h**eem	*to hit*
i	[i]	**i**n	*house*
ii	[i:]	km**ii**ng	*my name*
iiu	[i:w]	tkh**iiu**	*seven*
k	[k]	**k**eel	*goat*
kh	[k^h]	**kh**uung	*drum*
l	[l]	**l**oon	*hill*
m	[m]	**m**ei	*fire*
n	[n]	**n**ii	*day*
ng	[ŋ]	**ng**aa	*fish*
o	[ɔ]	n**o**n	*snout*
oo	[o:]	oon	*to call*
p	[p]	**p**uu	*grandfather/uncle*
ph	[p^h]	**ph**ul	*water pot*
r	[r]	**r**aal	*war*
s	[s], occasionally [ç]	**s**om	*ten*
t	[t]	**t**alu	*what*
th	[t^h]	**th**ung	*inside*
thl or ṭl*	[t͡ɬ]	**thl**aa	*moon/month*
tl	[t͡l]	**tl**oo	*do*
tx or ṭ*	[t̪s̪]	**tx**im	*half*
txh or ṭh*	[t̪s̪^{h}]	**txh**i	*to lead*
u	[u]	th**u**k	*come out*
uu	[u:]	n**uu**	*mother*
uui	[u:j]	**uui**	*dog*
v	[ʋ]	**v**ak	*pig*
y	[j]	**y**aan	*night*
'	[ʔ] or [ᵊ]	t**'**loo	*to take*

Table 3: Vowels and Consonants in Lamkang, adapted from Chelliah et al. (2023). *Grapheme versions without the combining dot diacritic are used in the current form of orthography

- learning_rate: 0.0001
- train_batch_size: 8
- eval_batch_size: 8
- seed: 42
- gradient_accumulation_steps: 2
- total_train_batch_size: 16
- optimizer: Use OptimizerNames.ADAMW_TORCH_FUSED with betas=(0.9,0.999) and epsilon=1e-08 and optimizer_args=No additional optimizer arguments

- lr_scheduler_type: linear
- lr_scheduler_warmup_steps: 300
- num_epochs: 100
- mixed_precision_training: Native AMP