End-to-End Speech Recognition for Endangered Languages of Nepal

Marieke Meelen
University of Cambridge
Cambridge, UK
mm986@cam.ac.uk

Alexander O'Neill
SOAS, University of London
London, UK
ao34@soas.ac.uk

Rolando Coto-Solano
Dartmouth College
New Hampshire, United States
rolando.a.coto.solano@dartmouth.edu

Abstract

This paper presents three experiments to test the most effective and efficient ASR pipeline to facilitate the documentation and preservation of endangered languages, which are often extremely low-resourced. With data from two languages in Nepal —Dzardzongke and Newar— we show that model improvements are different for different masses of data, and that transfer learning as well as a range of modifications (e.g. normalising amplitude and pitch) can be effective, but that a consistently-standardised orthography as NLP input and post-training dictionary corrections improve results even more.

1 Introduction

Of the 120+ distinct languages identified in the 2011 Nepali census, at least 60 are endangered due to socio-political unrest, globalisation and environmental challenges. The 2015 earthquake and the global pandemic have had devastating effects on the tourist industry, which formed the major source of income for the country. Long-lasting consequences include the increased migration away from the rural areas where many endangered languages are spoken towards Kathmandu and other areas where the Nepali language is dominant, as well as international destinations for education and employment. The loss of these languages also means the loss of unique cultural and religious identifiers. Given this, there is a clear need for methods and tools to preserve linguistic and cultural diversity.

A well-known challenge in language preservation, however, is the transcription bottleneck (Shi et al., 2021): transcribing one minute of audio requires at least an average of 40+ minutes (Durantin et al., 2017). The transcription process is furthermore severely hindered by the fact that many endangered languages do not have written traditions or standardised orthographies. While advanced automatic speech-recognition (ASR) tools are available, they are often ineffective for these extremely low-resource languages (Foley et al., 2018), due to the lack of good-quality training data.

In this paper we present results from three experiments aimed at creating ASR models for the endangered languages of Nepal: (a) Training models for two extremely low-resourced languages, Dzardzongke (South Mustang Tibetan) and Kathmandu Valley Newar, (b) testing the effectiveness of transfer learning for Dzardzongke from the related Standard Tibetan language, and (c) testing other techniques that are useful to enhance low-resource ASR such as sound and output manipulation, to measure their effectiveness on datasets of different sizes.

1.1 Languages

Dzardzongke or South Mustang Tibetan (SMT) is a severely endangered language spoken by maximum ca. 1200 people in a small number of villages in Mustang, Nepal. Most speakers of Dzardzongke are fluent in Nepali and Seke as well, and Dzardzongke is not used in writing or education, putting it in a very precarious situation. The difficult socio-economic situation in the aftermath of the 2015 earthquake and global pandemic is having a disastrous effect on the local language and unique pre-Buddhist Bon cultural tradition.

Newar, or Nepāl Bhāṣā, is a "definitely endangered" language (Moseley, 2010), with about 846,557 native speakers out of a population of about 1,321,933 ethnic Newars (Central Bureau of Statistics (Nepal), 2012). Newars live in 63 of the 77 districts of Nepal but are the indigenous inhabitants of the Kathmandu Valley, where they are centred and now make up a sizeable minority (Kansakar, 1999). While ethnic Newars mostly use Newar amongst themselves and in private, al-

Proceedings of the The Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 83–93

most all Newars use Nepali in public domains (Kansakar et al., 2011). Newar has been grouped into five geographical groupings, each including various dialects (Shakya, 2019). Our work in this paper utilised recordings from Lalitpur, Kritipur, and Kathmandu Newar, which, while they belong to the same geographical grouping (Kathmandu Valley Newar), are distinct dialects. In addition, we utilised historical recordings from Bhaktapur, which is from a distinct geographical grouping.

Speakers of both Dzardzongke and Newar are keen to preserve their language and cultural traditions and would therefore greatly benefit from the development of tools that can facilitate this preservation.

1.2 ASR for Low-Resource Languages

As mentioned in the previous section, the problem of the transcription bottleneck presents serious issues to language documentation. The ASR technology to perform this task is not new (Besacier et al., 2014), but it traditionally only achieved good results on large corpora. Recent years have seen work on end-to-end transcription of low-resource languages (Prud'hommeaux et al., 2021; Coto-Solano et al., 2022). These are made possible by the emergence of models that are pre-trained with acoustic data from other languages. These offer a robust acoustic model from their previous knowledge of multiple high-resource languages. Presently there is work beyond the high-resource languages, bootstrapping available data from low-resource languages to enhance both the acoustics and the textual output. Some techniques involve training with data from text-to-speech systems (Bartelds et al., 2023), and augmenting the data with other written sources such as dictionaries and word lists (Hjortnaes et al., 2020; Arkhangelskiy, 2021), as well as manipulating the transcription of the input (Coto-Solano, 2021).

One way to leverage data from other languages is to apply transfer learning. Transfer learning is a technique that uses knowledge from one language to improve the results of another with lower resources. It is a common technique in NLP fields like Machine Translation, where the model is trained on high-resource languages, and it is then fine-tuned on languages with fewer resources (Zoph et al., 2016; Kocmi and Bojar, 2018). This approach is useful when there are similarities between the source and target languages, be they genetic, typological or orthographic. Usually, a greater overlap in vocabulary between the high and low-resource languages leads to higher gains (Nguyen and Chiang, 2017; Dabre et al., 2017). However, this overlap is not necessary for models to benefit from transfer learning. In the case of ASR, models can pre-train on data from languages that are unrelated to the target, and even then the acoustic model section will see gains in performance (Bansal et al., 2019).

There are simple transformations that can help the model learn from the data. For example, researchers have found that manipulating acoustic characteristics like amplitude (Mitra et al., 2012) and pitch (Yadav and Pradhan, 2021) can lead to lower error rates.

2 Methodology

In this section we discuss our data collection and ASR pipeline, followed by a description of our experiments.

2.1 Data collection

The data for Dzardzongke was collected in August 2022 in a range of villages in Mustang.[1] We collected over 20 hours of interviews, conversations, as well as descriptions of rituals, traditional activities, and, finally read narratives in controlled environments. 251 minutes (4 hrs 11 minutes) are fully transcribed; over half of which containing read narratives by one near-native male speaker and the rest a mixture of conversational data from native speakers (2 male; 1 female, all 55+ years old). As Dzardzongke does not have any written history, we developed an orthography in collaboration with the local community. Unlike Standard Tibetan, this is a romanised script, which is not only more intuitive for native speakers who never learnt to read Tibetan, but also much more suited to the phonotactics of the language, yielding a straightforward mapping of sounds to graphemes. This enhances results of ASR models based on Wav2Vec2, as many of the languages in the pre-training set are written using the Roman alphabet. In total, the transcriptions contain 32,598 words in 5498 utterances, for an average of 5.9 words per utterance. There are 4664 unique words in the Dzardzongke transcriptions. The utterances are an average of 2.7 seconds long.

[1] All Dzardzongke audio-visual materials are available on ELAR http://hdl.handle.net/2196/70707494-ag7d-4hf2-ag77-fe21 (Meelen and Ramble, 2023).

The data for Newar comes from a combination of sources. 86 minutes of Kathmandu Newar were recorded in 2019 in a diaspora setting using read materials, the texts of which were later adapted for use with ASR.[2] 30 minutes of Bhaktapur Newar were used from historical recordings provided on a CC license by the CNRS's Pangloss project (Michailovsky and Sharma, 1968). The remaining data was collected in Nepal during fieldwork from August to November 2022. We collected 10 hours and 25 minutes of interviews, speeches, and spontaneous conversation, of which 185 minutes were fully transcribed and adapted for use with ASR. These were transcribed using the romanised IAST transliteration, which allows for one-to-one representation of and conversion to Devanagari, the script used for contemporary Newar. In addition to using data from four distinct dialects, this dataset includes data from two female speakers from Bhaktapur, yielding a combined total of 294 minutes (4 hrs 54 minutes) of transcribed data. The transcriptions contain 38,360 words in 4815 utterances, for an average of 8.0 words per utterance. The recordings are an average of 3.7 seconds long (1 second longer than our Dzardzongke recordings). There are 8038 unique words in the Newar transcriptions. Together, these factors mean that our Newar data would be a more significant challenge for training an ASR model.

2.2 ASR Training

We used Wav2Vec2 (Baevski et al., 2020) to train the models. First, we trained separate models for each language. We used different time partitions to measure the progress of the word error rate (WER) and the character error rate (CER) as the volume of data increases. We believe that these results could be valuable to other researchers in the area of extremely low-resource ASR, as they would give them an approximate idea how much data they would need to get the results they are aiming for. For both languages we randomly selected files until we reached partitions of [5, 10, 15, 30, 45, 60, 90, 120, 180] minutes. For Dzardzongke we also used a model trained on 251 minutes, the maximum amount of data available. For Newar, we also trained models of 240 and 294 minutes, the last one of which included all of the data available. For each of these, we randomly shuffled the dataset and distributed the available files into train/valid/test splits of 80%, 10%, 10%. We repeated this procedure ten times for the models without any input or output modifications. We trained on each of these and then retrieved the resulting model with the lowest WER validation values from the earliest possible checkpoint before overfitting. This model was then used to get the median CER and WER from the test set. The charts and tables below report the average values of the median over the 10 repetitions.

Wav2Vec2 uses multilingual quantisation to get better performance when transcribing sounds, and these might be to our advantage. The models presented here are "monolingual" in that the fine-tuning was done on only one of the languages (Dzardzongke or Newar), but the models are initialised from the highly multilingual XLSR-wav2vec2 base model, which includes data from 128 different languages. We used the instantiation in the Hugging Face (2024) libraries with their default parameters.[3]

2.3 Transfer Learning

Since Dzardzongke is related to Standard Spoken Tibetan, and the latter has a large amount of training data and an ASR model available, it is worth exploring the option of transfer learning from the higher-resourced language to the lower-resourced one. Although there are some distinct differences in vocabulary and morphosyntax, Standard Tibetan phonology is very similar to Dzardzongke. Unlike Dzardzongke, Standard Tibetan is widely spoken, not just in Tibet, but mainly in the Tibetan diaspora communities all over the world.

For the transfer-learning experiments, we trained our own small Standard Tibetan model based on 7 hours of training data, and also used a ready-made model based on 550 hours of training data, made available by OpenPecha.[4] Both models were later fine-tuned in the same way, by converting the Tibetan Unicode to Dzardzongke Romanised script output. These Standard Tibetan datasets contain a large variety of recordings, ranging from conversational data from media outlets (both TV and radio mainly based in Dharamsala, India) to Tibetan audiobooks and speeches from members of the Tibetan community.

[2]This is freely available on Zenodo https://zenodo.org/records/10611827.

[3]The hyperparameters, as well as the best performing models, can be downloaded from http://github.com/rolandocoto/nepali-asr.

[4]https://huggingface.co/datasets/openpecha/tibetan-voice-550

The test procedure is very similar to the one for the monolingual ASR models described above. We randomly selected audio files and put them in time partitions of [5, 10, 15, 30, 45, 60, 90, 120, 180, 251] minutes for Dzardzongke, and [5, 10, 15, 30, 45, 60, 90, 120, 180, 240, 294] minutes for Newar. We made five samplings for each of these time points (where we randomly selected from the entire pool of files for each language), and split them into 80%, 10% and 10% for the train/valid/test sets. From each training run we extracted the median CER and WER values for the best-performing model and calculated the average across the five different runs.

Standard Tibetan is written in a different script, however. Therefore, we also developed conversion rules to change the Standard Tibetan script to the newly-developed romanised Dzardzongke orthography. Since Newar is linguistically much further removed from Tibetan and there were no other datasets available for languages closer to Newar, we limited the transfer-learning experiments to Dzardzongke only for now.

2.4 Signal and output transformations

We performed several manipulations of the input wave files and the output transcriptions to improve our results.

Three of them included modifying the acoustic properties of the input audio files. In one subexperiment we normalised the amplitude (Mitra et al., 2012). We modified the audio files so that their peak would correspond to 70dB. These were then used to train a new monolingual model for each of the languages. The second modification was normalising the pitch, which has been observed to help with ASR in some populations, for example children (Shahnawazuddin et al., 2017). We changed the median pitch of all of the wave files to 151Hz. These new recordings were used to make another, separate model, so that we could compare these modifications to the performance with the unmodified wave files. For the third modification we included noise (Braun and Gamper, 2022), in particular pink noise, at a volume of 45dB. Pink noise has a more realistic and irregular distribution, compared to other types of synthetic noise, and could potentially make the system more robust in learning human speech. All of these modifications were carried out using the algorithms in the *Praat Vocal Toolkit* (Corretge, 2012).

The evaluation for these was performed in a similar way to the experiments above. Therefore, the results between the "no modification" condition and modifications are directly comparable. The error reporting is identical to the experiments above (average of the median WER and CER for all available test sets).

The final modification was the 'Dictionary word correction' performed on the output. When dealing with low-resource languages many non-words can be produced, which can ultimately undermine readability. In order to compensate for this, we introduced a series of simple modifications to the output. We used Norvig's (2021) unigram statistical spelling corrector but introduced one modification: if (i) the source and the ASR hypothesis transcription have the same number of words, and (ii) the word in source$_i$ is not the same as the word in hypothesis$_i$, then we will assume that the word hypothesis$_i$ is a spelling mistake and it will be changed to a different, existing word. This is meant to minimise the disruption on the output that standard statistical spell checking can introduce. We used the random shuffles from the monolingual models in section 2.2 and corrected their outputs here to make the spell checking results directly comparable to the "no modification" results.

3 Results

3.1 ASR Training

Table 1 shows the results of training monolingual models for each of the languages, when the models are trained for 30, 60 and 120 minutes of data. It also shows the models trained with the maximum amount of data for each language. Dzardzongke data achieved lower error rates despite having less data: WER=34 for 251 minutes, compared to WER=50 for the 294 minutes of Newar.

As Figure 1 shows, the character error rates drops relatively rapidly as the volume of data increases. The error for models trained on 5 minutes of data is CER=25 for Dzardzongke and CER=38 for Newar. Models trained on 60 minutes of data have half of this error (CER=11 for Dzardzongke and CER=18 for Newar), and subsequent models have smaller reductions: CER=8 and CER=12 for Dzardzongke and Newar respectively when training on all available data. The WER also follows a similar pattern, albeit with a slower reduction. When trained on 5 minutes, the Dzardzongke models have an average of WER=70, and the Newar

		CER				WER			
		30	60	120	Max	30	60	120	Max
Dzardzongke	No recording or output modifications	13	11	9	8	50	44	38	34
	Transfer from Tibetan (7 hrs)	13	10	8	**7**	50	42	35	33
	Transfer from Tibetan (550 hrs)	12	9	8	**7**	49	39	35	33
Dzardzongke	Normalise amplitude	11	9	8	**7**	48	41	37	33
	Normalise pitch	13	10	9	**7**	52	43	39	33
	Pink noise	14	13	9	8	50	46	43	33
	Word correction	14	11	9	8	46	41	34	**32**
Newar	No recording or output modifications	25	18	16	**12**	74	63	59	**50**
	Normalise amplitude	19	16	16	**12**	64	57	54	**50**
	Normalise pitch	22	17	18	14	67	67	60	**50**
	Pink noise	20	17	16	13	67	67	57	**50**
	Word correction	40	20	17	14	77	61	55	**50**

Table 1: Average error rates for ASR models of Dzardzongke (max 251 mins) and Newar (max 294 mins).

models WER=94. Models trained on 60 minutes have approximately 65% of the error (WER=44 and WER=63). The errors are halved by the time the models are trained with all the available data (WER=34 for Dzardzongke and WER=50 for Newar).

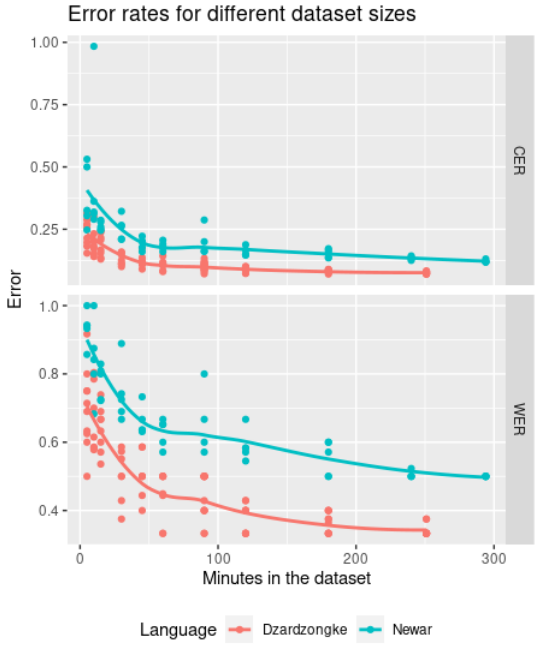

Figure 1: Character and word error rates for ASR training in Dzardzongke and Newar, by the minutes in the combined train-eval-test sets.

It is noteworthy that there is a wider gap between the languages in the word error rate. When using all available data the difference in character error between the two languages is ΔCER=4, but the difference in word error rate is ΔWER=16. This might be because of the architecture of Wav2Vec2. Given that it uses quantisation based on phones from numerous languages, it has more information about the sounds of human languages in a straightforward romanised representation, which is closer to the orthography that was developed especially for Dzardzongke than the non-standardised transcriptions found in the diverse Newar varieties.

3.2 Transfer Learning results

Table 1 also shows the performance of the transfer learning experiments, where Standard Tibetan models were used as a basis to enhance the results for the related Dzardzongke language. When trained on all available data, there is only a small gain: the WER is reduced by one unit for both of the transfer models (WER=34 for no transfer; WER=33 for 7 or 550 hours of Tibetan). The CER is also reduced by one (CER=8 for no transfer; CER=7 for 7 or 550 hours of Tibetan).

Figure 2 shows the difference in error rates when trained with different amounts of data. The gains from transfer learning are greater when the model has fewer minutes of the target language available. For example, when training on 60 minutes of data, the model transferring from 550 hours of Tibetan has a WER=39, 5 units lower than the WER for the model without transfer (WER=44). The model transferring from 7 hours of Tibetan has more modest gains (ΔWER=2 points; WER=42), but it also improves results. Even when you only have two hours of data the gains are still present: the transfer models had WER=35 compared to WER=38 without transfer. As mentioned above, these gains begin to disappear as the data in the target language increases.

3.3 Signal and output transformations

The second and third sections of Table 1 show the average results for the signal and output transformations performed on the Dzardzongke and Newar

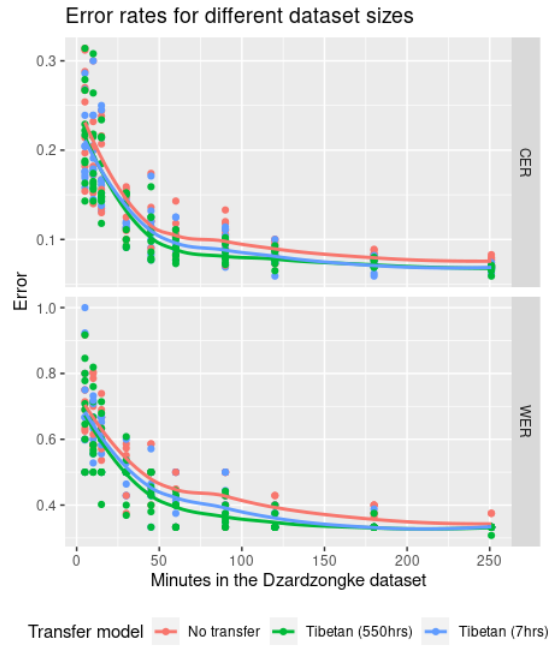

Figure 2: Transfer learning from two Tibetan models.

models. When training on all the data, applying the word correction to the output of the Dzardzongke monolingual (i.e. non-transfer-learning) model provides improvements in word error rate (WER=32). As for CER, normalising the amplitude and the pitch provided a small improvement for Dzardzongke (WER=7, compared to WER=8 without modifications). None of the modifications improved the results of the Newar model trained with all its data; they all reached a WER=50, and normalising the amplitude produced the same CER results as making no modifications (CER=12).

Figure 3 shows the result of the transformations done on datasets of different sizes. In the case of Dzardzongke, there is virtually no difference between the conditions when it comes to CER performance. Normalising the amplitude produces gains of approximately ΔCER=2 (e.g. at 30 and 60 minutes of total data), but normalising the pitch does not lead to improvements. Adding (pink) background noise and correcting the words can make the CER worse.

Some of the modifications do have a positive impact on the WER of Dzardzongke. For example, applying the word corrections to the 30 minute datasets improves the results by ΔWER=4 (46, compared to WER=50 for the non-corrected version). These gains diminish as data increases, but they are still present. When the dataset has 60 minutes, the gain is ΔWER=3 (41, compared to WER=44 for non-corrected), and when the dataset has 2 hours of audio, the gain is ΔWER=4 (34, compared to WER=38 for non-corrected). Normalising the amplitude also produced improvements (e.g. ΔWER=1∼2), but normalising the pitch and adding noise can produce increases in error rates.

The modifications produce more improvements in the Newar data. As for the CER, all the modifications of the audio improved the error rates to some degree, with normalisation in amplitude being the one that reduced the error the most (ΔWER=2∼6). Normalising the amplitude also produced gains in the WER. When training on 30 minutes, there was a gain of ΔWER=10 (64, compared to 74 for non-modified audio). The improvements from amplitude normalisation became smaller when training on 60 minutes (ΔWER=6) and on two hours of data (ΔWER=5), and they finally disappear when training on the maximum amount of data. Adding pink noise also leads to some improvements (ΔCER=0∼5, ΔWER=2∼7), but normalising the pitch can lead to increases in error rates. Unlike Dzardzongke, applying word corrections does not consistently improve the CER and WER of Newar. When training on 30 minutes of data, the error increases (ΔCER=-15, ΔWER=-3), but when training on 60 and 90 minutes of data, there are some improvements in the WER (ΔWER=2∼4), but not in the CER (ΔCER=-1∼-3).

In summary, normalising the amplitude of the signal seems to uniformly decrease the error rates, while applying word corrections can lead to WER reduction for Dzardzongke in particular.

3.4 Transcription results

The first part of Table 2 shows four transcription results for Dzardzongke. Example (1) contains a number of phonetic difficulties, like the similarity between the velar and palatal nasal in front of high vowels (*nyí* vs *ngi*) and the difference between high and low tone (indicated by an acute accent, e.g. *léparak* vs *leparak*). Finally, it shows that rare personal names can be difficult to transcribe. These difficulties can be remedied by adding more monolingual data, as shown by the improved error rates comparing the 5 vs 251 min models (ΔCER=4; ΔWER=24). Example (2) shows similar improvements in a more challenging utterance from a conversation in a noisy environment, whose WER can be improved even further using the spell checking on the output. This does not work for the highly

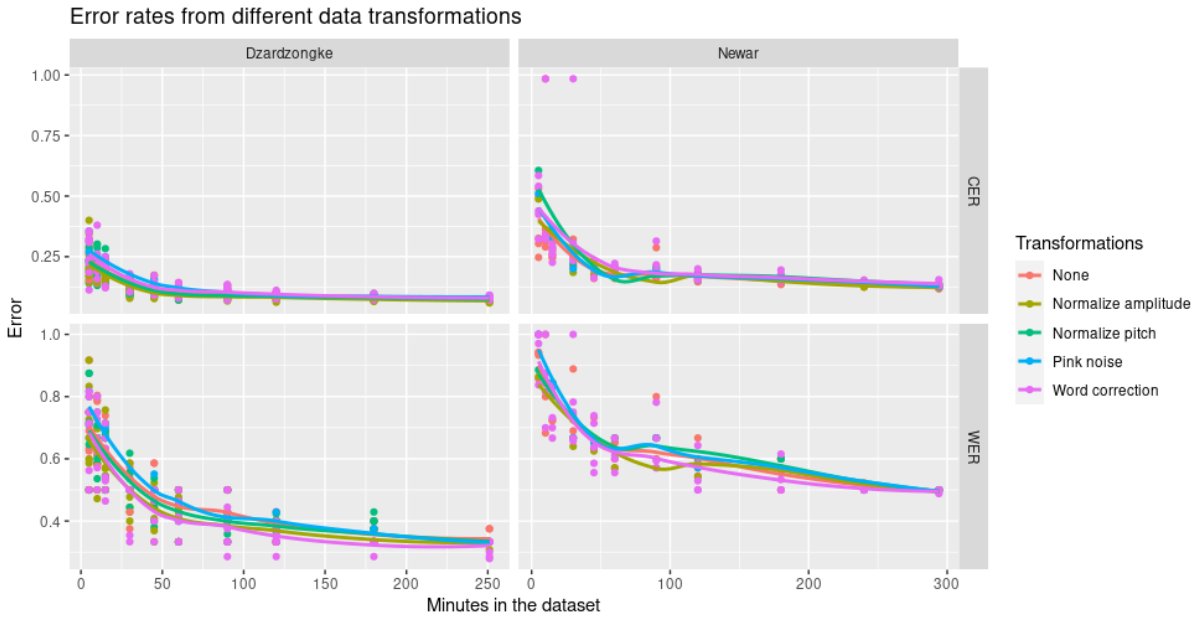

Figure 3: Transformations of input and output of Dzardzongke and Newar models.

	Dzardzongke controlled near-native narrative		
1. [smt-041-296]	"When 2 (more) came, Ódrong-Gepo arrived at the end"		
Target transcription	*nyí ongna ódrong gepo katsa ru dzangi léparak*	CER	WER
5 mins	*ngi o nga odrong gepo katsa ru dzangi leparak*	13	62
251 mins	*nyi onga ódrong gepo katsa ru dzangi leparak*	**7**	**38**
	Dzardzongke native conversation in noisy environment		
2. [smt-005-896]	"Then all of a sudden, having gone outside,"		
Target transcription	*da japtsowe phita la sori*	CER	WER
5 mins	*ta jzapdi phital sori óo*	48	100
251 mins	*da zaptsi phitala sori*	**20**	60
251 mins + Dict	*da lapti phita sori*	32	**40**
	Dzardzongke very bad CER and WER		
3. [smt-005-587]	"...the girl I like ..."		
Target transcription	*... ngi sempa ... theken bomo*	CER	WER
251 mins + transfer	*de sempi ta nyikure nyi sempa te bomo*	110	150
	Dzardzongke very good CER and WER		
4. [smt-037-390]	"Look, I only have 150 rupees at the moment"	CER	WER
251 mins + transfer	*nga la danda ale gya dang ngápcu mana me* = target	**0**	**0**
	Lalitpur Newar		
5. [ltp-016-2526]	"...a preacher of the Dharma..."		
Target	*dharmabhānaka*	CER	WER
294 mins	*dharma bhānaka*	8	200
	Kritipur Newar		
6. [VM-VM2-157]	"First of all,"		
Target	*dakale nhāpā*	CER	WER
294 mins	*dakal nhāpāṃ*	17	100
294 mins + Dict	*dakale nhāpāṃ*	**8**	**50**
	Bhaktapur Newar very bad CER and WER		
7. [HD-HD-260]	"Now, that's not the case. You..."		
Target	*āḥ thva athe makhu chiṃ*	CER	WER
294 mins	*aāmaka thvāṅā ānikāṃ thātheyake naypiṃ yaṅ āḥ thaḥre makhu chiṃ*	178	160
	Lalitpur Newar very good CER and WER		
8. [ltp-016-3930]	"You deigned to say to me, 'O son of good family,..."'	CER	WER
294 mins	*vasapolapiṃsaṃ jita dhayā bijyāta he kulaputra* = target	**0**	**0**

Table 2: ASR results from various experiments for Dzardzongke and different varieties of Newar

infrequent *japtsowe* 'all of a sudden', whose orthography exceptionally differs significantly from pronunciation [japtsi] (almost captured by the model).

Finally, (3) and (4) respectively show representative examples of very bad and very good transcriptions. The wave file for example (3) actually con-

tains noise at the start and middle of the utterance, leading to the ellipsis for missed words in the target transcription (which were automatically filtered out as punctuation during training). The best Dzardzongke model (max + 550 hrs transfer) actually does a very reasonable job, but the error rates are very high due to the incomplete original transcription. To improve overall results, utterances with incomplete transcriptions due to noise etc. should therefore be filtered out before the training to avoid skewing the overall error rates. Example (4) on the other hand is from a narrative in a controlled, quiet environment and is one of many such examples for which the best model yields perfect transcriptions. Although many of these zero-error transcriptions come from these narrative, controlled recordings, the model is already robust enough to generalise beyond this one speaker as shown by results from the noisy, conversational recordings like (2).

Table 2 also highlights the success and difficulty we encountered with Newar and some examples of why our WER is misleading when evaluating this model's quality. Example (5), for instance, whose target was *dharmabhānaka* was recognised as *dharma bhānaka*. While the CER=8 was good, it had an extra word than the source, resulting in WER=200. However, inconsistent spacing in Newar orthography means the result is legitimate; thus, we can qualitatively assign this a true CER and WER of 0. This issue consistently resulted in a high WER for Newar when in fact the result was qualitatively acceptable.

Word separation and a unigram-based probabilistic calculation for the spell checking meant that our corrected outputs were less optimal than we would have liked. However, Kritipur Newar (6) is an example of a success of spelling correction, where the target was *dakale nhāpā* was initially recognised incorrectly as *dakal nhāpāṃ*, but the automatic correction changed this to *dakale nhāpāṃ*. While the resulting WER=50 is expected, as the second word in the source was *nhāpā*, again, the flexibility of Newar orthography means that *nhāpāṃ* is both a standard and acceptable variant of *nhāpā*. Therefore, we could qualitatively assign this example a true CER and WER of 0.

Example (7) is taken from a public performance recording, where the target only shows the speaker's speech, but the ASR model also identified the speech of an audience member. As with the incomplete Dzardzongke transcription of example (3), this utterance should either be removed or completed before training.

In (8), finally, we see an example of how this ASR model could perfectly recognise complicated and relatively lengthy speech. If one considers that the first two Newar examples are also qualitatively perfect, these examples demonstrate that with the careful selection of training data, one can develop optimal ASR models for low-resource languages without too much difficulty.

4 Discussion

From the results of all three experiments it becomes apparent that modifications are most useful up to around 90 minutes of 'monolingual' transcriptions. Transfer learning in particular proved more effective at this stage than sound modifications, although the size of the Standard Tibetan datasets mattered less than expected.[5]

The Newar dataset exhibits a broad heterogeneity, encompassing a wide range of sources, whereas the Dzardzongke data originates from a more specific geographical area with more data from one speaker, and, on average, shorter utterances, which could explain the higher Newar WER of 50 (vs Dzardzongke 32). Additionally, the Newar collection primarily features literary works, including readings of literature, theatrical performances, and discourses on religious or literary subjects that do not generalise well to more casual conversations that are also part of the same dataset.

Post-training corrections based on probabilistic spell checking from existing monolingual transcriptions is marginally effective for improving WER in Dzardzongke, but would be more effective especially for recordings on new topics if a more comprehensive corpus were available.

For Newar, the lack of standardised, romanised spelling leads to higher word (but not character) error rates, but as shown in the previous section, these are not necessarily representative of actual qualitative errors in transcription.

For both languages, as well as the Standard Tibetan datasets, an in-depth analysis of transcriptions results reveals the importance of a well-balanced, varied dataset where incomplete transcriptions are filtered out to avoid artificially high error rates that make the models worse. Although it is tempting with any low-resource language to

[5] More information on the content and accuracy of Standard Tibetan transcriptions was not available.

utilise as many transcribed utterances as possible, those with too much noise or interference are clearly creating more problems later on.

5 Conclusion

Our main goal was to present a first test of the most effective and efficient ASR pipeline to facilitate the documentation and preservation of endangered languages, which are often extremely low-resourced. For both Dzardzongke and Newar, model improvements are different for different masses of data, which helps to guide those who have to start transcriptions from scratch.

We tested different modification techniques to see which would be most effective for small-size datasets and carefully evaluated and discussed the results. Directions for future research include experiments with transfer learning for Newar and further modifications and corrections once word lists in standardised orthographies have been created.

Limitations

There are some limitations in the current datasets upon which the models were trained. First, they are still of limited size and the Newar set in particular is very heterogenous as it contains samples from four different varieties. The Dzardzongke dataset on the other hand is less robust since half of the data consists of recordings of narratives by one near-native speaker in a quiet, controlled environment. For both datasets, most speakers are old and there are very few women.

Additionally, the training took a large amount of processing time, and this might be prohibitive for many teams and communities. The models were trained using Nvidia Tesla K80 GPUs from Dartmouth's Research Computing, and training all the models took approximately 3972 hours of computing time. This was done in an HPC infrastructure, with 5∼7 processes running in parallel. With this set up, the training took approximately one month. The inference per se does not consume so many resources, but a user would still need a GPU to actually get a transcription. While this can be done online with a number of free alternatives, the cost could be prohibitive for communities who wish to implement these transcription systems offline.

Ethics Statement

Ethics approval was obtained prior to data collection from the University of Cambridge.

Acknowledgements

We would like to thank Esukhia/Monlam AI for sharing their smaller Standard Tibetan dataset with us at an early stage allowing us to start training our transfer models. We also thank Charles Ramble, Nyima Drandul and Birat Raj Bajracharya for support with the Dzardzongke and Newar transcriptions. We also gratefully acknowledge funding for various parts of this project from the Endangered Language Documentation Programme (ELDP - G114548), the Cambridge Centre for Digital Humanities Incubator Grant 2023 and the Arts and Humanities Research Council (AHRC - AH/V011235/1). Finally, we want to thank Jianjun Hua, Elijah Gagne and the personnel at Dartmouth Research Computing for their assistance in setting up the experiments.

References

Timofey Arkhangelskiy. 2021. Low-resource ASR with an augmented language model. In *Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages*, pages 40–46, Syktyvkar, Russia (Online). Association for Computational Linguistics.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 58–68, Minneapolis, Minnesota. Association for Computational Linguistics.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation. *arXiv preprint arXiv:2305.10951*.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic Speech Recognition for Under-Resourced Languages: A Survey. *Speech Commun.*, 56:85–100.

Sebastian Braun and Hannes Gamper. 2022. Effect of noise suppression losses on speech distortion and ASR performance. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 996–1000. IEEE.

Central Bureau of Statistics (Nepal). 2012. *National Population and Housing Census 2011: National Report*. Central Bureau of Statistics, Kathmandu.

Ramon Corretge. 2012. Praat vocal toolkit. *Barcelona, Spain: Praat. Retrieved from http://praatvocaltoolkit.com*.

Rolando Coto-Solano. 2021. Explicit tone transcription improves ASR performance in extremely low-resource languages: A Case Study in Bribri. Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas (pp. 173–184). Association for Computational Linguistics.

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, and Isaac Feldman. 2022. Development of Automatic Speech Recognition for the Documentation of Cook Islands Māori. Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 3872-3882). https://aclanthology.org/2022.lrec-1.412.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 282–286. The National University (Phillippines).

G. Durantin, B. Foley, N. Evans, and J. Wiles. 2017. Transcription survey. *Paper presented at the Australian Linguistic Society Annual Conference*.

B. Foley, J. T. Arnold, R. Coto-Solano, G. Durantin, T. M. Ellison, D. van Esch, and J. Wiles. 2018. Building speech recognition systems for language documentation: The CoEDL endangered language pipeline and inference system (ELPIS). *SLTU*, pages 205–209.

Nils Hjortnaes, Timofey Arkhangelskiy, Niko Partanen, Michael Rießler, and Francis Tyers. 2020. Improving the language model for low-resource ASR with online text corpora. 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 336–341, Marseille, France. European Language Resources association.

Hugging Face. 2024. Transformer Documentation: XLSR-Wav2Vec2. `https://huggingface.co/docs/transformers/model_doc/xlsr_wav2vec2`.

Tej R. Kansakar. 1999. The Sociology of the Newar Language. *Newar Vijñāna*, 2:17–27. INBSS, Portland.

Tej R. Kansakar, Nirmal Man Tuladhar, Omkareshwor Shrestha, Shobha Kumari Mahato, Narayan Gautam, Sulochana Sapkota, and Kishore Rai. 2011. A Sociolinguistic Survey of Newar/Nepal Bhasa. A Report submitted to the Linguistic Survey of Nepal.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Marieke Meelen and Charles Ramble. 2023. An Audio-Visual Archive of Dzardzongke (South Mustang Tibetan). Endangered Language Archive.

Boyd Michailovsky and Ramapati Raj Sharma. 1968. The Sahu Hari Das has many troubles. Audio recording, Pangloss: A CNRS Project.

Vikramjit Mitra, Horacio Franco, Martin Graciarena, and Arindam Mandal. 2012. Normalized amplitude modulation features for large vocabulary noise-robust speech recognition. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4117–4120. IEEE.

Christopher Moseley, editor. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. UNESCO Publishing, Paris.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Peter Norvig. 2021. How to Write a Spelling Corrector. `https://norvig.com/spell-correct.html`.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic Speech Recognition for Supporting Endangered Language Documentation. *Language Documentation & Conservation*, 15:491–513.

Syed Shahnawazuddin, KT Deepak, Gayadhar Pradhan, and Rohit Sinha. 2017. Enhancing noise and pitch robustness of children's ASR. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5225–5229. IEEE.

Daya Ratna Shakya. 2019. Agreement vs Non-Agreement: Gradual Development of Inflectional Pattern, Assessment drawn from Ten Dialects of Nepal Bhasa. *Nevāḥ Prajñā*, 2(3):41–80.

J. Shi, J. D. Amith, R. Castillo García, E. G. Sierra, K. Duh, and S. Watanabe. 2021. Leveraging end-to-end ASR for endangered language documentation: An empirical study on Yolóxochitl Mixtec. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, page 1134–1145.

Ishwar Chandra Yadav and Gayadhar Pradhan. 2021. Pitch and noise normalized acoustic feature for children's ASR. *Digital Signal Processing*, 109:102922.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.