Exploring ASR fine-tuning on limited domain-specific data for low-resource languages

Mak, Franco
Voice Computing Research Group, CSIR Next Generation Enterprises and Institutions Cluster, Pretoria, South Africa
francomak@protonmail.com

Govender, Avashna
Voice Computing Research Group, CSIR Next Generation Enterprises and Institutions Cluster, Pretoria, South Africa
agovender1@csir.co.za

Badenhorst, Jaco
Voice Computing Research Group, CSIR Next Generation Enterprises and Institutions Cluster, Pretoria, South Africa
jacbadenhorst@gmail.com

Abstract

The majority of South Africa's eleven languages are low-resourced, posing a major challenge to Automatic Speech Recognition (ASR) development. Modern ASR systems require an extensive amount of data that is extremely difficult to find for low-resourced languages. In addition, available speech and text corpora for these languages predominantly revolve around government, political and biblical content. Consequently, this hinders the ability of ASR systems developed for these languages to perform well especially when evaluating data outside of these domains. To alleviate this problem, the Icefall Kaldi II toolkit introduced new transformer model scripts, facilitating the adaptation of pre-trained models using limited adaptation data. In this paper, we explored the technique of using pre-trained ASR models in a domain where more data is available (government data) and adapted it to an entirely different domain with limited data (broadcast news data). The objective was to assess whether such techniques can surpass the accuracy of prior ASR models developed for these languages. Our results showed that the Conformer connectionist temporal classification (CTC) model obtained lower word error rates by a large margin in comparison to previous TDNN-F models evaluated on the same datasets. This research signifies a step forward in mitigating data scarcity challenges and enhancing ASR performance for low-resourced languages in South Africa.

Keywords: Automatic speech recognition, fine-tuning, low-resource languages, data harvesting, broadcast news data

1 Introduction

Automatic Speech Recognition, a model that converts spoken language into text, is becoming increasingly integrated into our daily lives. ASR nowadays plays a pivotal role in powering voice assistants, transcription and dictation services and most importantly has become a transformative tool that can enable communities to communicate with devices in several languages. However, the development of ASR has been largely skewed towards high-resourced languages that often cater for Western cultures, creating an accessibility gap for communities communicating in low-resourced languages. In South Africa, this divide is pronounced where 10 of the 12 official languages of South Africa are low-resourced and face a critical shortage of labelled data for training effective ASR models Badenhorst & de Wet (2022). In addition, existing speech and text corpora for these languages are primarily centred around government, political and biblical content. This limited coverage in content leads to high error rates when attempting to transcribe content in domains outside of this scope. Our research aims to tackle this challenge by exploring methodologies where we can leverage existing data using adaptation and fine-tuning techniques to increase accuracy on out-of-domain content which will then contribute to creating new domain specific labelled datasets. For example, radio broadcast data is a largely untapped resource in South Africa which offers a dynamic and ever-growing pool of audio content in most of South Africa's languages. Leveraging this resource can pave the way for enhanced

ASR models specifically catered to radio broadcast news.

In this work, we have focused on 3 South African languages namely: Afrikaans (Afr), isiZulu (Zul) and Sesotho (Sot). Each of these languages belong to different language families in South Africa. isiZulu is a Nguni language and Sesotho is one of the Sotho languages. Since there are distinct acoustic and linguistic differences between these languages, evaluating these languages will provide meaningful insights on whether the techniques explored in this paper would yield high quality transcriptions when evaluated on the remaining low-resourced languages in South Africa that generally fall under these two broad language families.

2 Previous Work

The feasibility of harvesting radio broadcast speech data for the development of ASR systems for South African languages was first evaluated in Badenhorst & de Wet (2021). In this work, a semi-automatic data harvesting procedure was proposed. Factorised time-delay neural network (TDNN-F) models were used to generate phone-level transcriptions of speech data harvested from different domains. The results showed that when evaluating on speech data from a new domain for Afrikaans, the phone error rates (PERs) of approximately 20% was measured. At these PERs, follow-up experiments confirmed a potential word transcription rate within the thirties for news data (Badenhorst & de Wet 2023). Transcription error rates in other languages however measured higher and so for the purpose of creating correctly-annotated speech data the first publication already concluded that better acoustic modelling techniques need to be sought.

The work presented in Badenhorst & de Wet (2023) further investigated the possibility of creating diverse speech resources from unannotated radio broadcast data using a refined version of the semi-automatic data harvesting procedure proposed in Badenhorst & de Wet (2021). The work focused on the adaptation of ASR models on two domains within broadcast data, namely news bulletins and radio dramas. Baseline models were trained using NCHLT (Barnard et al. 2014) data which mostly includes short read speech utterances. Adapting these baseline models to news and drama data, as expected, resulted in more transcription errors. The main reasons for this is that news data differs significantly from NCHLT speech data in that it contained much longer utterances. The radio drama data differed significantly to NCHLT data in speaking style as drama data corresponds more to conversational speech rather than read speech found in the NCHLT data. In addition, the drama data contained background noise. However, improvement in error rates were achieved by adapting acoustic models with less than 10 hours of manually annotated data from the same domain, for the speaking styles and acoustic conditions that are not represented in any of the existing speech corpora.

3 Experimental Design

3.1 Data and Preparation

3.1.1 Training Data

The National Centre for Human Language Technology (NCHLT) speech corpus is the largest publicly available corpus available that includes text and speech data comprising of approximately 55 hours of short-form audio segments from approximately 200 speakers in all 11 official written languages of South Africa (Barnard et al. 2014). This dataset was used to train the baseline models in this work, referred to as pre-trained models. This dataset is comprised of mono-channel audio files with a sampling frequency of 16 kHz and 16-bit rate. This speech data contains clear, read speech prompts with minimal background noise. Table 1 summarizes the audio statistics of the NCHLT speech corpus.

3.1.2 Adaptation Data

The broadcast news dataset used as adaptation data for fine-tuning in this work was obtained as part of an on-going data harvesting project to record new audio data from South African radio stations.

Table 1: Summary statistics for NCHLT data

	Afr	Zul	Sot
Total number of speakers	210	210	210
Average audio duration (s)	3.10	4.50	3.50
Training set duration (h)	53.70	52.20	53.80
Training set segments	63131	41871	54817
Test set duration (h)	2.66	4.02	2.54
Test set segments	3002	2802	2722

The corresponding transcriptions for this data were automatically transcribed using the data harvesting procedure proposed in Badenhorst & de Wet (2023), resulting in datasets comprising approximately 150 hours (for each language) of transcribed audio for six South African languages with coverage of news bulletins. The six languages include: Afrikaans, isiZulu, Sesotho, isiXhosa, Tshivenda, and Sesotho sa Leboa.

A limited subset of this automatically generated transcriptions was manually verified to correct for any transcription errors present which resulted in 15 hours of verified labelled audio data per language. During manual verification each transcription was marked for inaccuracies within square brackets. For example, all unintelligible words were marked with "[?]", all words from other languages were marked with "['foreign'+transcription word]" and all partial or incomplete words were marked as word fragments. After manual verification, an analysis was performed on the output transcriptions to compile a list of all the transcribed inaccuracies present in the dataset. The results showed that the radio news readers frequently use English words in the broadcast, otherwise known as code-switching. Since the model development in this work focused only on language specific models comprising of a single language source, only the audio segments for which the corresponding transcriptions contained no inaccuracies were used. In this way, the adaptation data did not contain any foreign or unintelligible words. Table 2 summarizes the audio statistics for the broadcast news adaptation dataset.

Table 2: Summary statistics for broadcast news data after removing all segments containing inaccuracies

	Afr	Zul	Sot
No. of news bulletins	377	433	430
Average audio duration (s)	6.90	7.00	7.20
Training set duration (h)	16.34	8.70	8.80
Total available segments	8528	4737	4684

3.2 Model Selection

Kaldi is an open source speech recognition toolkit for ASR and signal processing. In Badenhorst & de Wet (2019), TDNN-F models were trained following recipes from the Kaldi toolkit (Povey et al. 2011). The Kaldi toolkit has since been updated to Kaldi 2 (K2) along with newer model recipes found in an open source repository called Icefall. A wide variety of model recipes for streaming and non-streaming ASR is available in Icefall. In this work, three Icefall recipes were chosen for initial exploration, which include: the Conformer CTC, pruned transducer stateless and TDNN LSTM CTC recipes. The Conformer CTC model is a transformer model (Gulati et al. 2020) combined with a convolution module and uses a CTC (or connectionist temporal classification) for loss and decoding (Graves et al. 2006). The pruned transducer stateless is a recurrent neural network transducer (RNN-T) model with a faster loss computation involving pruning bounds of RNN-T recursion (Kuang et al. 2022). The third model is the Time Delay Neural Network (TDNN) with Long Short-Term Memory (LSTM) utilising CTC loss. For each of these recipes, a baseline isiZulu model was trained using the NCHLT isiZulu dataset. Across the 3 implementations, the Conformer CTC model produced the lowest word error rates (WERs) and was therefore the model of choice.

Icefall provides three versions of the Conformer CTC model. According to documentation found in the Icefall repository, the differences between each variation is as follows:

1. **Conformer CTC**: The original Conformer model that was used for training the isiZulu

baseline model.

2. **Conformer CTC 2**: According to a Results.md document found in the Icefall Github repository, the authors have improved the original Conformer CTC by implementing a reworked CTC attention model that they claim trains faster, offers better stability and slightly better WERs.
3. **Conformer CTC 3**: A Conformer that implements a streaming mode that supports symbol delay penalty. The delay penalty is a method that lowers streaming latency while having minimal impact on recognition accuracy (Kang et al. 2022).

Models for each of the new versions of the Conformer CTC model were trained using the same NCHLT isiZulu dataset. Across the three models compared, the Conformer CTC 2 performed the best. The same model selection process was repeated for the NCHLT Afrikaans dataset, and the same trend as isiZulu emerged. For this reason, the Conformer CTC 2 model was chosen for the work that followed.

3.3 Model Training

Table 3: Model hyper parameters during training

Parameter	Value
Training	
Initial learning rate	0.003
Batch sizes	5000
Conformer model	
Feature dimension	80
Subsampling factor	4
Encoder layers	12
Encoder dimensions	512
Feedforward dimension	2048
CTC loss	
Beam size	10
Reduction	sum
Use double scores	True

The baseline Conformer CTC 2 model was trained for all 3 languages on the respective NCHLT language datasets: Afrikaans, isiZulu and Sesotho. Table 3 summarizes the model hyper-parameters used during training. The same parameters were used for all 3 baseline models. The models were trained for 30 epochs, and then each epoch was evaluated to determine the phone-error-rate (PER). The PER was the main metric used to evaluate performance, because phoneme-level evaluation gives a more precise report of the frequency of substitution, omission, and addition errors the model made compared to word-level evaluation. To gauge transcription performance, subsequent word error rates (WERs) are only computed at selected PERs using language models borrowed from previous text-based refinement work done in Badenhorst & de Wet (2023). These language models were 3-gram ARPA models built from the CTexT NCHLT Text Corpora (Puttkammer et al. 2014) of the relevant language.

To commence fine-tuning, the aim was to firstly obtain a pre-trained NCHLT baseline model that was most optimised for the broadcast news domain, so the test set that was used to evaluate the performance of each baseline epoch was one hour of broadcast news data. The epoch with the lowest PER was then selected. During fine-tuning, these pre-trained baseline models was further trained using the broadcast news adaptation data until the model converged again.

3.3.1 Test data

The news adaptation dataset given in Table 2 is small. Reserving a portion of the already limited dataset for use as a test set might not serve as a good indicator of the model's actual performance. Therefore, 5-fold cross validation was used to evaluate the performance of the Conformer CTC 2 model after fine-tuning. The news adaptation data was randomly split into 5 equal subsets. The NCHLT baseline model was fine-tuned through additional training on four of the subsets, with the fifth subset serving as the test set. The model was evaluated, and then the fine-tuning procedure was restarted with the same NCHLT baseline but with a different sub-

set used as the test set. Repeat this process until each subset has served as the test set. The average recognition result over the five models served as the overall performance metric of the model.

A separate, smaller dataset of broadcast news data was also available for model evaluation. This is an older news dataset that was used in Badenhorst & de Wet (2023) to evaluate the performance of previous TDNN-F models. The source of this dataset is the same as the news adaptation data described in Table 2, but there is no overlap. In this paper, this older test set will only be used to compare the results of the new Conformer CTC 2 models against previously obtained TDNN-F results. Table 4 describes the audio statistics of this older news test set.

Table 4: Summary statistics for older set of broadcast news data used in previous work (Badenhorst & de Wet 2023)

	Afr	Zul	Sot
Test set duration (min)	473.95	30.54	43.19
Average audio dur (s)	7.38	40.73	51.83
Number of segments	3852	45	50

4 Results

4.1 Baseline Models

To create a pre-trained NCHLT baseline model the Conformer CTC 2 recipe was applied to NCHLT training data. Table 5 shows the best recognition performance obtained for each baseline model. All word-error-rates and phone-error-rates in this paper were obtained from a 1-best decode, which extracts the best path from the decoding lattice as the decoding result.

During training it was observed that the amount of epochs required for NCHLT models to converge on broadcast news recognition in each language differed. The model for Afrikaans optimised after only 6 epochs, while isiZulu and Sesotho both required 17 epochs before the model started to over-fit towards NCHLT.

Table 5: Best phone-error-rate and word-error-rate for each baseline model evaluated on one hour of broadcast news test set

Language	Epoch	PER	WER
Afrikaans	6	29.58	53.98
isiZulu	17	27.39	52.21
Sesotho	17	71.02	76.66

4.2 Fine-tuned Models

The best performing epoch from pre-training shown in Table 5 was chosen for further training by fine-tuning it on the broadcast news adaptation dataset for another 25 epochs. As mentioned in Section 3.3.1, a 5-fold cross validation strategy was employed to evaluate the performance of the model after fine-tuning on the broadcast news domain. Table 6 shows the best performance of each fine-tuned model evaluated on its respective test subset.

From the results presented in Table 6, it is evident that PER and WER improved after fine-tuning the baseline model on the broadcast news adaptation data. The standard deviation for PER and WER from each of the five subsets is small, which indicates that the model performance is consistent and stable.

In Table 7 below, we compare the PER and WER of the TDNN-F model trained in previous work (Badenhorst & de Wet 2023) against the fine-tuned Conformer CTC 2 model. The test set used here is an older news dataset described in Table 4 above. The Conformer CTC 2 models from Subset 2 of the 5-fold cross validation set from Table 6 was used for evaluation, because they achieved low PER for the three languages.

4.3 Impact of domain specific data

Experimental results above showed that fine-tuning a baseline NCHLT model on additional news adaptation data resulted in improved recognition when evaluated on broadcast news data. However, ASR models generally gain recognition improvements

Table 6: Lowest phone-error-rates and word-error-rates of each of the five cross validation subsets of broadcast news adaptation data

	Afrikaans		isiZulu		Sesotho	
Subset	PER	WER	PER	WER	PER	WER
Subset 1	13.20	20.0	17.77	45.95	34.66	42.35
Subset 2	12.26	19.96	16.00	45.3	36.22	43.03
Subset 3	12.86	20.08	16.12	45.91	36.28	44.03
Subset 4	12.82	20.48	16.59	45.18	36.20	43.35
Subset 5	12.63	19.98	16.19	45.84	36.16	44.02
Mean	12.75	20.10	16.53	45.64	35.90	43.36
Standard deviation	0.31	0.19	0.65	0.33	0.62	0.63
Variance	0.09	0.04	0.42	0.11	0.39	0.40

Table 7: Comparison of Word-error-rates between TDNN-F taken from Badenhorst & de Wet (2023) and fine-tuned Conformer CTC model from Subset 2 of the cross-validation model set

Language	TDNN-F		Conformer CTC 2	
	PER	WER	PER	WER
Afrikaans	20.98	37.42	18.91	29.74
isiZulu	38.81	69.24	25.57	58.35
Sesotho	73.50	56.89	43.66	45.01

when given additional training data. To compare the performance gained from fine-tuning a baseline model on domain-specific data relative to using more out-of-domain data of the same quantity, an additional experiment using 15 hours of the Afrikaans broadcast news data (Table 2) was conducted.

To compare the impact of in-domain versus out-of-domain data, 15 hours of Afrikaans NCHLT training data given in Table 1 was randomly selected for use as out-of-domain adaptation data. A new baseline model was then trained on the remaining 38 hours of Afrikaans NCHLT data for 20 epochs. Again, the epoch with the best performance when evaluated on one hour of an independent broadcast news test set was chosen as the starting baseline model. This model was duplicated and separately fine-tuned on the two sets of adaptation data for 20 epochs. The best recognition results for both models when evaluated on the broadcast news test set is given in Table 8 below. The epoch number that resulted in the best performance is also included. When fine-tuning, the baseline model was considered as epoch zero.

Table 8: Comparison of error rates after fine-tuning on either in-domain news data or out-of-domain NCHLT data

Model version	Epoch	PER	WER
Baseline NCHLT	7	28.72	52.41
Adapt NCHLT	4	28.67	51.94
Adapt News	12	16.77	22.92

5 Discussion

In this work, the aim was to explore various models to determine whether it is possible to reduce transcription errors on an out-of-domain test set. Our exploration commenced by first training 3 types of Conformer CTC models using NCHLT data which contains content that mostly lends itself to government data whilst the data that we wish to accurately transcribe or test on contains news content. The results of these baseline NCHLT models when evaluated on out-of-domain broadcast news data is shown in Table 5. The lowest WER results of the baseline models when tested on one hour of broadcast news was around 50% for Afrikaans and isiZulu, and 77% for Sesotho. These high error rates indicate that the model is over-fitting to the content

contained in NCHLT during pre-training, and thus performed poorly on unseen data outside of this domain. It was also observed that for Afrikaans the model converged the fastest whilst for isiZulu and Sesotho the model took longer to converge.

We hypothesise that there could be a number of reasons for the differences. First and foremost there are morphological differences between the languages themselves. Niesler et al. (2005) performed an analysis of the phone sets and phone transcriptions between four South African languages (Afrikaans, English, isiZulu and isiXhosa), and found that isiZulu was substantially more phonetically complex and diverse than Afrikaans. They concluded that speech recognition may be expected to be intrinsically more difficult for isiZulu and isiXhosa compared to Afrikaans and English, which could be why the Conformer CTC 2 model was able to learn the acoustic properties of Afrikaans more easily.

From previous work it was also evident that acoustic motivations exist for ASR performance differences given the NCHLT speech corpora. The TDNN-F 1d model analysis in Badenhorst & de Wet (2019) showed that Afrikaans NCHLT test PERs were approximately double that of isiZulu and Sesotho. In fact, the Sesotho PER was the second highest of all the languages. Therefore, acoustically, the NCHLT training data for isiZulu and Sesotho did not seem to be as effective even for in-domain NCHLT ASR compared to Afrikaans. It is not entirely clear why the error rates were so different, but in addition as shown in Table 1, isiZulu and Sesotho NCHLT training segments have longer average duration than Afrikaans. The lengthier segments could for instance include additional acoustic variability such as various background effects.

To improve transcription accuracy we investigated the ability of the model to generalise to other domains by exposing the model to a limited amount of adaptation data. Thus, we fine-tuned the baseline models by training the model further on the broadcast news adaptation data from which our test data is derived. Due to the limited size of this adaptation data, a 5-fold cross-validation method was utilised to evaluate the overall performance of the Conformer CTC 2 model. The results in Table 6 showed that improvements in the recognition metrics with low standard deviations were achieved across all three languages. The model from Subset 2 was selected for further evaluation on an older broadcast news test set in Table 7, and showed improved recognition performance over previous TDNN-F model results from Badenhorst & de Wet (2023).

An alternative explanation for the improved recognition metrics after fine-tuning could simply be that more training data was introduced to the baseline model. To investigate the impact of domain-specific data, the experiment described in Section 4.3 was performed. Creating and adapting a 38 hour baseline NCHLT model with 15 hours of in-domain and out-of-domain data showed that introducing the additional 15 hours of out-of-domain NCHLT data had a very small impact on the recognition metrics of the news test set. However, introducing 15 hours of in-domain news data successfully reduced the recognition metrics by half.

6 Conclusion

Developing ASR models for low-resource languages in South Africa is faced with challenges arising from the scarcity of data. It is even more challenging to develop ASR systems that are better fit-for-purpose in domains outside the ones in which existing data presides. The work presented in this paper explored methodologies to address this challenge, specifically investigating the viability of fine-tuning baseline models—initially trained on existing data for our low-resource languages—through limited adaptation data in targeted domains, with the aim of enhancing transcription accuracy.

In this pursuit, radio broadcast news data emerged as a valuable audio resource. The acquisition of correctly annotated transcriptions for this data proves to be a laborious and resource-intensive undertaking. By being able to generate accurate transcriptions of radio broadcast data, we anticipate a significantly positive impact on data harvesting endeav-

ours in the future. As this, in turn, contributes to the development of automatically transcribed labelled data that can be used towards continuous advancements in ASR in the country.

New modelling techniques presented in this work such as Conformer CTC 2 together with additional fine-tuning on limited adaptation data has demonstrated promising outcomes in reducing transcription errors on radio broadcast data. This marks a meaningful stride towards successfully generating more accurate labelled corpora in our local languages, signifying progress in the ongoing development of localising ASR to our South African languages.

Acknowledgements

The manual verification of automatically transcribed broadcast news data was enabled by funding received from the Department of Science and Innovation through the South African Centre for Digital Language Resources (SADiLaR), https://www.sadilar.org/. We are also indebted to the Centre for High Performance Computing (CHPC) in Cape Town for providing computing resources.

References

Badenhorst, J. & de Wet, F. (2019), 'The usefulness of imperfect speech data for ASR development in low-resource languages', *Information* **10**(9).

Badenhorst, J. & de Wet, F. (2021), 'Investigating the feasibility of harvesting broadcast speech data to develop resources for South African languages', *Journal of the Digital Humanities Association of Southern Africa* **3**(03).

Badenhorst, J. & de Wet, F. (2022), 'NCHLT Auxiliary speech data for ASR technology development in South Africa', *Data in Brief* p. 107860.

Badenhorst, J. & de Wet, F. (2023), 'Gauging the accuracy of automatic speech data harvesting in five under-resourced languages', *Journal of the Digital Humanities Association of Southern Africa* **4**(02).

Barnard, E., Davel, M., van Heerden, C., Wet, F. & Badenhorst, J. (2014), 'The NCHLT speech corpus of the South African languages', *SLTU 2014* pp. 194–200.

Graves, A., Fernández, S., Gomez, F. & Schmidhuber, J. (2006), 'Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks', *Proceedings of the 23rd International Conference on Machine Learning* p. 369–376.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y. & Pang, R. (2020), 'Conformer: Convolution-augmented Transformer for Speech Recognition', *arXiv preprint arXiv:2005.08100* .

Kang, W., Yao, Z., Kuang, F., Guo, L., Yang, X., lin, L., Żelasko, P. & Povey, D. (2022), 'Delay-penalized transducer for low-latency streaming ASR', *arXiv preprint arXiv:2211.00490* .

Kuang, F., Guo, L., Kang, W., Lin, L., Luo, M., Yao, Z. & Povey, D. (2022), 'Pruned RNN-T for fast, memory-efficient ASR training', *arXiv preprint arXiv:2206.13236* .

Niesler, T., Louw, P. & Roux, J. (2005), 'Phonetic analysis of Afrikaans, English, Xhosa and Zulu using South African speech databases', *Southern African Linguistics and Applied Language Studies* **23**, 459–474.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. & Vesely, K. (2011), 'The Kaldi Speech Recognition Toolkit', *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* .

Puttkammer, M., Schlemmer, M., Pienaar, W. & Bekker, R. (2014), 'NCHLT Afrikaans text corpora'.