BLSTM harvesting of auxiliary NCHLT speech data

Jaco Badenhorst, Laura Martinus & Febe de Wet
Human Language Technology Research Group
CSIR Meraka Institute, Pretoria, South Africa
jbadenhorst@csir.co.za, lmartinus@csir.co.za, febe.dewet@gmail.com

***Abstract*—Since the release of the National Centre for Human Language Technology (NCHLT) Speech corpus, very few additional resources for automatic speech recognition (ASR) system development have been created for South Africa's eleven official languages. The NCHLT corpus contained a curated but limited subset of the collected data. In this study the auxiliary data that was not included in the released corpus was processed with the aim to improve the acoustic modelling of the NCHLT data. Recent advances in ASR modelling that incorporate deep learning approaches require even more data than previous techniques. Sophisticated neural models seem to accommodate the variability between related acoustic units better and are capable of exploiting speech resources containing more training examples. Our results show that time delay neural networks (TDNN) combined with bi-directional long short-term memory (BLSTM) models are effective, significantly reducing error rates across all languages with just 56 hours of training data. In addition, a cross-corpus evaluation of an Afrikaans system trained on the original NCHLT data plus harvested auxiliary data shows further improvements on this baseline.**

***Index Terms*—NCHLT corpora, speech data, under resourced languages, automatic speech recognition, Bidirectional Long Short Term Memory, Kaldi**

I. Introduction

In September 2009, the Department of Arts and Culture (DAC) of the South African government put out a call for proposals for the development of speech and text resources for the country's eleven official languages. These resources were to be delivered to the National Centre for Human Language Technology (NCHLT) with the aim to advance the development of human language technology (HLT) in South African languages. The aim of the NCHLT Speech project was to create speech resources for the development of text-to-speech (TTS) and automatic speech recognition (ASR) systems, for the eleven official languages in South Africa. The project was carried out by a research group at the CSIR's Meraka Institute and the corpora were delivered to the DAC in 2013.

The ASR data that was made publicly available after the completion of the project constitute a subset of the data that was collected. The released data was selected from the total pool of collected data to satisfy the project specifications [1]. Very few speech resources have been developed for the country's official languages since the NCHLT project. The aim of the study reported on here was therefore to investigate the potential value of the data that was collected but not included in the NCHLT speech corpus.

ASR systems rely on large volumes of transcribed data from which acoustic models can be derived. The variation that is expected to occur in speech data should be represented in the training data to build representative models. Training data should therefore be diverse, containing examples that represent as much as possible of the variation typically observed in speech signals.

The initial version of the data collection tool resulted in a high repetition of a limited number of prompts. It would not be good practice to include too many examples of the same utterances in a training set, so the repeated prompts were mostly excluded from the NCHLT Speech corpus. However, it is often said that *"there is no data like more data"* and some of the more recent acoustic modelling techniques do indeed seem to be capable of using just about any training data to improve modelling accuracy. The aim of this study was therefore to determine how much of the un-released NCHLT data is potentially useful and whether simply having *"more"* data could improve the performance of ASR for South Africa's official languages.

II. Background

While it may be true that *"there is no data like more data"* it also holds that models trained on bad data will produce poor results: *"garbage in, garbage out."* We thus wanted to identify utterances that were not suitable for model development and exclude those from the pool of auxiliary data.

The NCHLT data collection protocol included a number of crude checks to identify corrupt and empty recordings. In our current investigation we also eliminated prompts that could not be aligned with the phone string expected to be produced when they are pronounced. In addition, we used a phone-based dynamic programming (PDP) scoring technique [2] to rank recordings according to the degree of acoustic match between the expected and produced prompts.

The ASR results that were published with the first release of the NCHLT Speech corpus were obtained using both the HTK [3] and Kaldi [4] toolkits. At the time the best results were obtained using the Kaldi implementation of Subspace Gaussian Mixture Models (SGMMs). A more recent study on one of the languages (Xho) suggests that substantial gains over the initial baseline can be achieved with Deep Neural Net (DNN) approaches [5]. Since the study published in [5], it has been shown that time delay neural network (TDNN) [6], [7] and long short-term memory (LSTM) acoustic models outperform systems based on DNNs [8], [9].

Further improvements were reported for bi-directional LSTMs (BLSTMs) that process input data in both time di-

rections using two separate hidden layers. BLSTMs allow the preservation of both past and future context information [10]. The interleaving of temporal convolution and BLSTM layers has been shown to model future temporal context effectively [11]. Furthermore, for BLSTM training on limited data (10-50 hours), as much as 5 layers of depth seem to be better than 3 layers. For training sets approaching 100 hours of data even better performance can be obtained using 6 deep layers [12]. TDNN-BLSTM acoustic models also yielded the best results in this study.

In the remainder of this paper we report on the extent of the repetition in the NCHLT auxiliary data as well as the techniques that were used to identify potentially useful recordings. In addition we present new baseline results for the NCHLT Speech data and investigate the utility of the auxiliary data by conducting initial ASR experiments using newly harvested data.

III. Extended corpora

As was mentioned in Section I, not all the data that was collected during the project was included in the final NCHLT Speech corpus because the initial recordings only represented a limited vocabulary. A second phase of data collection was initiated with updated data collection tools. As a result, two data sets were collected for a number of languages: one set with many examples of a limited vocabulary and one set with fewer examples of a more diverse vocabulary.

A. Speech data

After data collection was completed, three data sets were created using a progressive data selection strategy to construct the final deliverable [1]:

1) *NCHLT-raw*
   The total set of usable data collected after all empty and otherwise unusable recordings were discarded. This includes multiple sessions of some speakers and multiple examples of some prompts.
2) *NCHLT-baseline*
   A subset of *NCHLT-raw* representing approximately 200 unique speakers per language and more than 200 utterances per speaker. Recordings from the more diverse second batch of data were given preference in cases where speakers participated in both data collection campaigns.
3) *NCHLT-clean*
   A subset of *NCHLT-baseline* constituting the final deliverable of ± 56 hours of speech data for all 11 official languages.

All three of these data sets contain prompted speech. Prompts were derived from the biggest text corpus that was available for each language [13]. A text selection algorithm was used to optimise vocabulary coverage using the most frequently observed n-grams for each language.

A mobile data collection tool was subsequently used to record the prompts while they were read out by participants [14]. These recordings were not manually annotated. Instead, a confidence scoring technique was used to identify recordings that did not match their associated transcriptions. Poor matches usually occur as a result of reading errors, high levels of background noise, hesitations, etc.

The recordings with the best confidence scores (well-matched with their associated transcriptions) and that contributed most to lexical diversity were included in the final corpora [1]. These criteria were used to select an equal amount of data (± 56 hours of speech) for all 11 languages. As a result, data of a sufficiently good acoustic quality was excluded from the final corpora for some languages. We refer the data in *NCHLT-baseline* not included in *NCHLT-clean* as *Aux1*. It should be borne in mind that *Aux1* contains utterances produced by the same speakers as in the *NCHLT-clean* data set. *Aux2* includes all utterances from *NCHLT-raw* that are not in *NCHLT-baseline*.

Table I presents the initial number of recordings (*init*) in the *Aux1* and *Aux2* data sets for each language[1]. The *failed* column in the table shows how many utterances in each data set failed the alignment process described in Section V. The percentage value in the last row of the table indicates that more than 90% of both the data sets could be aligned and could therefore be considered for harvesting. This corresponded to 780.57 and 640.70 hours of audio in *Aux1* and *Aux2* respectively.

TABLE I
TOTAL NUMBER OF AUXILIARY RECORDINGS (*Aux1* & *Aux2*), NUMBER OF FAILED PHONE ALIGNMENTS (*failed*) AND DURATION [H] OF ADDITIONAL DATA PER LANGUAGE.

Lang	**Aux 1**			**Aux 2**		
	init	**failed**	**dur**	**init**	**failed**	**dur**
Afr	54 117	2 451	42.68	47 290	356	39.14
Eng	42 958	952	29.78	54 719	628	38.92
Nbl	37 669	3 224	42.56	100 402	4 202	120.07
Nso	65 224	2 259	64.89	53 318	947	51.80
Sot	74 457	5 858	73.86	47 938	700	43.51
Ssw	67 410	7 172	78.41	136 422	9 490	167.00
Tsn	69 655	1 953	70.15	35 156	356	36.98
Tso	71 311	3 781	83.67	2 316	1 489	0.65
Ven	82 895	4 886	93.69	44 666	1 220	54.94
Xho	90 560	8 739	102.95	53 269	2 549	54.95
Zul	77 833	3 471	97.93	30 319	327	32.74
Total	734 089	6.1%	780.57	605 815	3.7%	640.70

B. Unique and repeated prompts

A first analysis of unique and repeated prompts in the *NCHLT-clean* data was conducted shortly after the corpus was released [15]. Tables II and III provide type and token counts for the prompts in the *NCHLT-clean*, *Aux1* and *Aux2* data sets. The values in the NCHLT TRN *Type* column correspond to the number of unique prompts in the NCHLT training set. The counts for prompt types that occur in the test set but *n*ot in the training set are listed in the NCHLT TST *Type* column.

[1]Three character ISO codes are used to refer to the 11 official languages in all the tables in this paper: Afrikaans (Afr), English (Eng), isiNdebele (Nbl), Sepedi (Nso), Sesotho (Sot), Siswati (Ssw), Setswana (Tsn), Xitsonga (Tso), Tshivenda (Ven), isiXhosa (Xho), isiZulu (Zul).

TABLE II
TYPE & TOKEN COUNTS FOR PROMPTS *only* IN NCHLT_TRN & *only* IN NCHLT_TST. AUX1 AND AUX2: TYPE AND TOKEN COUNTS FOR PROMPTS REPEATED IN AUXILIARY DATA.

	NCHLT TRN		Aux1		Aux2		NCHLT TST		Aux1		Aux2	
Language	**Ty**	**To**	**Ty**	**To**	**Ty**	**To**	**Ty**	**To**	**Ty**	**To**	**Ty**	**To**
Afr	9 482	39 589	8 268	30 494	996	29 224	44	44	44	299	0	0
Eng	6 509	33 595	5 724	22 425	1 934	17 095	95	106	86	301	9	14
Nbl	9 967	29 416	7 056	20 639	9 964	63 833	599	632	403	724	196	278
Nso	14 247	45 803	12 415	41 453	6 787	34 699	223	291	194	556	28	61
Sot	9 414	34 010	8 273	42 105	3 561	23 714	122	122	122	485	0	0
Ssw	9 781	28 472	9 097	33 662	9 781	79 138	160	164	158	687	2	2
Tsn	13 230	40 994	11 206	41 768	1 588	28 533	407	443	160	309	32	32
Tso	10 517	34 265	10 144	42 177	646	659	173	179	173	911	0	0
Ven	14 188	37 456	13 085	49 008	6 738	34 037	436	439	434	1 527	0	0
Xho	11 416	26 713	9 470	43 812	2 190	11 651	511	511	201	818	0	0
Zul	7 580	19 585	7 220	34 330	1 191	9 760	277	299	276	1 377	0	0

TABLE III
TYPE & TOKEN COUNTS FOR PROMPTS IN BOTH NCHLT_TRN *and* NCHLT_TST. AUX1 AND AUX2: TYPE & TOKEN COUNTS FOR PROMPTS REPEATED IN AUXILIARY DATA. NEW UNIQUE: TYPE & TOKEN COUNTS FOR NEW PROMPTS IN AUX1 AND AUX2.

	NCHLT TRN_TST		Aux1		Aux2		New Unique Aux1		New Unique Aux2	
Language	**Ty**	**To**	**Ty**	**To**	**Ty**	**To**	**Ty**	**To**	**Ty**	**To**
Afr	2 463	23 328	2 318	14 565	1 089	16 697	1 244	6 378	80	1 013
Eng	2 804	40 673	2 627	16 065	2 455	35 894	583	3 215	195	1 088
Nbl	2 269	9 393	1 696	4 366	2 269	16 326	2 450	8 716	2 716	15763
Nso	2 082	10 258	1 783	5 818	1 015	6 466	3 513	15 138	1 969	11 145
Sot	1 726	20 600	1 680	15 111	814	18 998	2 507	10 898	937	4 526
Ssw	2 292	11 898	2 189	9 219	2 292	2 546	3 442	16 670	3 448	22 376
Tsn	868	14 137	682	8 316	528	3 454	4 596	17 309	223	2 781
Tso	2 476	10 626	2 427	8 505	6	6	2 706	15 937	148	162
Ven	2 331	8 979	2 193	7 834	1 041	3 732	3 987	19 640	1 641	5 677
Xho	1 057	16 419	1 500	15 081	1 024	36 792	5 636	22 110	490	2 277
Zul	1 814	21 844	1 772	22 915	1 040	19 296	2 321	15 740	262	936

NCHLT TRN_TST types correspond to unique prompts that occur in both the training and the test set[2].

The *Aux1* and *Aux2* columns indicate how many of these *Type*s also occur in the auxiliary data. The type and token counts for the unique prompts that occur only in the auxiliary data are provided in the last four columns of Table III. The values in these tables indicate that the auxiliary data mostly contains repetitions of prompts that are already in the *NCHLT-clean* corpus.

C. Phone representations

The data analysis in this study required phone level transcriptions to process utterances. Text pre-processing was required to prepare the transcriptions for pronunciation extraction. All text was converted to lowercase and unwanted symbols (not within the list of graphemes for a particular language) were removed. Since numerous additional words occurred in the auxiliary data, the existing NCHLT pronunciation dictionaries had to be extended before the data could be processed.

During the NCHLT project, a set of grapheme-to-phoneme (G2P) rules were derived from the so-called *NCHLT-inlang* dictionaries [1]. These rules were used to predict pronunciations for the new words. No explicit procedure was followed to identify out-of-language words, but for certain languages the in-language G2P rules did not contain rules for particular graphemes or the punctuation mark used to indicate an apostrophe in English (Eng). For these words the Eng G2P rules were used to generate pronunciations and the phones were mapped to similar sounds using the in-language phone set.

Eng was the only language for which a different procedure was followed. G2P rules trained on a version of the Oxford Advanced Learner's dictionary, adapted to South African Eng using manually developed phoneme-to-phoneme rules were used for the analysis of the Eng data [16].

IV. NCHLT-CLEAN BASELINE REVISITED

This section presents a more recent baseline ASR recipe for the *NCHLT-clean* corpora. The *train*, *development* and *test* sets defined in [1] were used throughout.

A. ASR systems

We built phone recognition systems following the same Kaldi recipes used in [5] to create Triphone, SGMM and DNN-HMM hybrid models. TDNN-BLSTM models were also implemented by adapting the Kaldi Wall Street Journal (WSJ) example recipe [4].

The TDNN-BLSTM acoustic models were trained using 40-dimensional high-resolution MFCC features. The high-

[2]Type and token counts for the NCHLT DEV set are not included in the table. On average, the development sets contain around 3 000 prompt tokens.

resolution MFCCs were derived from speed[3] and volume[4] perturbed data.

Since the TDNN-BLSTM recipe required high-resolution MFCC features, a standard MFCC front-end with a 25ms Hamming window and a 10ms shift between frames (16 kHz sampling frequency) was employed to train all the other models. Mean and variance normalisation operations, applied on a per speaker basis, followed the extraction of 13 cepstra which included C0. Delta and double delta coefficients were added. These features were used to estimate 3-state left-to-right HMM triphone models, incorporating linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT) training and speaker adaptive training (SAT). SGMM training followed. The Kaldi *nnet2* setup was used to train DNN-HMM hybrid models keeping the same parameter settings as in [5], [18].

The TDNN-BLSTM network was generated with the *nnet3* Kaldi setup. We replaced the *nnet3* component graph with a similar TDNN-BLSTM structure obtained from a Switchboard chain model example recipe. This graph contained 1 standard, 3 time-delay and 3 BLSTM layers. For all layers the cell-dimension was set to 1024. The BLSTM forward and backward layers implemented delays of -3 and 3 respectively, setting recurrent and non-recurrent parameter dimensions of 256 and a decay time of 20. The remaining training parameters provided in the WSJ recipe were used without adjustment except for minibatch sizes (set to either 32 and 16) to enable 70 parallel CPU training jobs on a node with an 80 GB memory constraint.

B. Phone recognition measurement

A position independent phone configuration was used to convert the training transcriptions to a phone level representation. During system evaluation, this arrangement seamlessly converts the standard Kaldi word error rate (WER) measurement to a phone error rate (PER). Estimations of all PERs used speech phone labels only, ignoring any silence labels. Recognition employed a flat ARPA language model consisting of equiprobable 1-grams.

The best ratio between acoustic and language model contributions was determined by varying the language-scale parameter (integer values in the range of 1-20) during scoring. The acoustic-scale parameter was set to the default value of 0.1 and the best language-scale parameter was chosen using the development data sets previously defined for *NCHLT-clean* [1]. The selected language-scale parameters were subsequently used during data harvesting and to gauge recognition performance.

C. Results

Table IV shows the development (dev) and test set (tst) PER results of SGMM, DNN, and the baseline TDNN-BLSTM models for each language. As in [1] the number of phone labels (#Phns) provide an indication of the label complexity.

[3]Using factors of 0.9, 1.0 and 1.1 [17].

[4]Choosing a random factor between 0.125 and 2.

TABLE IV
PERs FOR SGMM, DNN AND TDNN-BLSTM BASELINE SYSTEMS PER LANGUAGE

		SGMM		DNN		BLSTM	
Lang	**#Phns**	**dev**	**tst**	**dev**	**tst**	**dev**	**tst**
Afr	37	11.31	12.73	8.50	9.63	5.89	6.64
Eng	44	18.68	18.42	14.06	13.73	7.69	7.24
Nbl	49	20.62	21.43	16.59	18.03	10.04	10.77
Nso	44	14.54	14.91	12.05	12.60	9.29	9.64
Sot	39	21.42	21.47	16.21	16.48	11.44	11.92
Ssw	39	17.12	16.65	13.39	13.71	9.17	8.70
Tsn	34	15.44	14.36	11.98	11.14	8.24	7.17
Tso	55	15.38	13.58	11.12	11.16	7.10	6.67
Ven	39	15.14	15.60	12.47	13.20	8.61	9.10
Xho	53	20.00	20.18	16.46	15.72	11.20	11.25
Zul	46	16.66	18.10	13.33	15.25	10.18	10.72

The results for the SGMM models are similar to the phone recognition result obtained with HTK in the 2014 study [1].

In general, PERs improved with more sophisticated models. The table shows that substantially lower PERs were obtained using TDNN-BLSTM models in comparison with SGMMs and DNNs. In fact, in most cases the TDNN-BLSTM PERs were almost half the corresponding SGMM values.

V. DATA HARVESTING

The purpose of automatic data harvesting is to detect acoustically compromised recordings so that they are not used as train or test data during system development. Section V-A describes the mechanism we used to rank recordings in terms of acoustic quality. Quantifying the acoustic variability in the data enabled the selection process described in Section V-B.

A. Acoustic ranking

For each language, we processed all of the *Aux1* and *Aux2* data using the improved baseline acoustic models described in Section IV. The harvesting procedure required each utterance to be decoded twice. Firstly, the standard free phone decode implementing an ergodic phone loop generated a sequence of phone labels, purely based on the acoustics. Next we used the supplied Kaldi functionality to compute training alignments from lattices for *nnet3* models. This algorithm generates a decoding graph for a single fixed sequence of phone labels, which directly corresponds to the reference transcription. In the event that the acoustics are not a good match for the forced sequence of phone labels, this constraint can result in the decode operation exiting without producing any output. Such unsuccessful decodes served as a first selection criterion to filter out large transcription errors.

As explained in Section II, PDP scoring matches the free phone decode and forced phone label sequences. It is possible to adjust the PDP algorithm using a cost matrix so that string edit operations (*substitution*, *deletion* and *insertion*) contribute differently for the various phone labels [19]. We opted to use a flat phone matrix where the contributions of the edit operations are the same for every phone label. *Insertions* and *deletions* contributed half as much to the score as *substitutions* and correctly recognised labels.

B. Data selection

This section reports on our attempt to improve ASR performance for two languages adding additional data from *Aux1*. To select the suitable subsets of additional training data, we estimated local PERs for 400 utterances at a time.

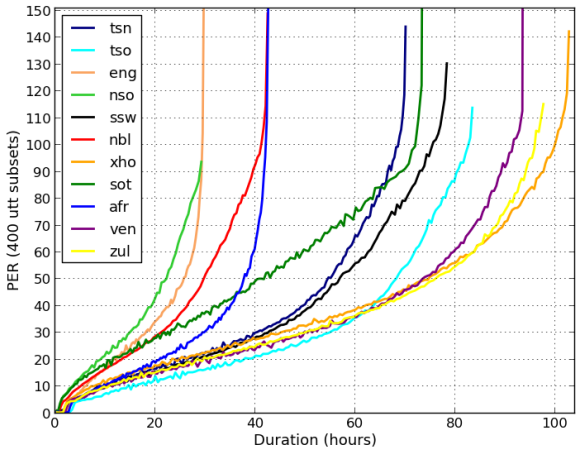

Fig. 1. Local phone error rates (PERs) for 400 utterance subsets of the *Aux1* data.

Figure 1 depicts graphs of the local PERs. These values were computed for non-overlapping subsets of utterances, ordered according to PDP scores. Figure 1 reveals a large range of PER scores for different subsets of utterances. PERs of higher than 100% can occur due to, for example, runaway insertions during free phone recognition. At an operating point of 50% PER, more than 20 hours and for some languages even more than 60 hours of additional data can be selected.

We decided to use a conservative estimate of 30% PER. Applying this threshold, we selected 29.8 hours of Afr and 18.9 hours of Eng data. The selected data also contained additional test data (for the same speakers as the *NCHLT-clean* development and test sets). Excluding these utterances resulted in 27.8 hours of additional data for Afr and 17.7 hours for Eng.

C. Selection validation

Our evaluation included cross-corpus test data to determine whether overtraining on the NCHLT corpus occurred. Section V-C1 introduces these data sets and explains the creation of the required phone representations.

1) Cross-corpus data: During the RCRL project [20] 330 Afr radio news bulletins that were broadcast between 2001 and 2004 on the RSG radio station were purchased from the SABC. The data was transcribed to create a corpus of around 27 hours of speech data. For the validation purposes in this study a previously selected 7.9 hour evaluation set containing 28 speakers was used.

The 20 hour South African broadcast news (SABN) corpus was compiled using broadcasts of one of South Africa's main radio stations, SAFM. The news bulletins were recorded between 1996 and 2006 and contain a mix of newsreader speech, interviews and crossings to reporters at remote locations [21]. We compiled a 3.5 hour subset of speech from 26 speakers as validation data.

To obtain the phone sequences from the RSG and SABN orthographies, we implemented the same procedure as for the NCHLT Afr and Eng systems. After text pre-processing, G2P rules were applied to generate pronunciations for new words.

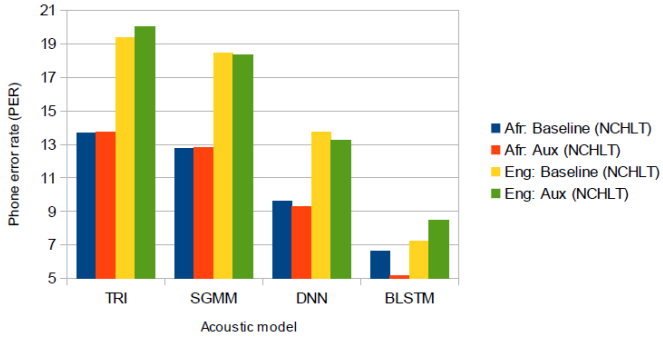

Fig. 2. Comparing PERs of all acoustic models on NCHLT test data.

2) Validation experiments: Figure 2 depicts the recognition results before and after data augmentation for *NCHLT-clean* test data. Overall the Afr systems produced lower PERs than the Eng systems. For Afr lower PERs were obtained for DNN-based systems, especially for the TDNN-BLSTM models (the latter dropping from 6.64% to 5.14%). The results for the Eng systems did not follow the same trend. While the DNN acoustic models produced a small gain (13.23% compared to 13.73%), the augmented system yielded a higher PER with the TDNN-BLSTM model (8.46% compared to 7.24%).

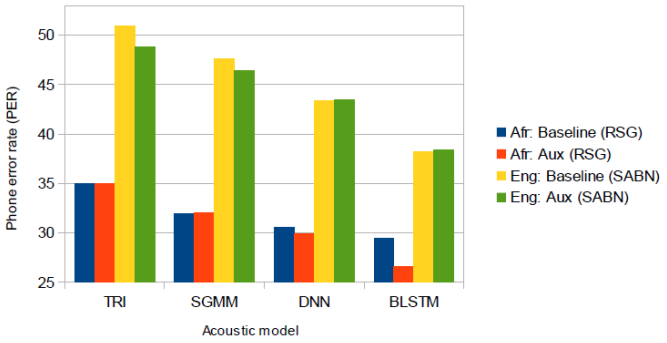

Fig. 3. Comparing PERs of all acoustic models on RSG and SABN test data.

Cross-corpus recognition results are illustrated in Figure 3. Decoding RSG data, the Afr system produced trends similar to those observed for the NCHLT test data. Again, the DNN and TDNN-BLSTM systems yielded performance gains with PERs dropping from 29.49% to 26.60%. Interestingly, the augmented Eng system produces slightly better results for triphone and SGMMs employing fewer parameters.

VI. Conclusion

This paper introduced a new ASR baseline for the entire NCHLT Speech corpus. Even with the available 56 hour corpora, deep learning architectures consistently produced substantial performance gains, lowering PERs considerably. The paper also described the large portion of previously unreleased auxiliary NCHLT data. Acoustic confidence scores could be obtained for close to 90% of the auxiliary data using the TDNN-BLSTM baseline ASR to perform data harvesting. Two sets of additional audio data (*Aux1* and *Aux2*) with a total duration of more than 1400 hours were compiled. Since the speaker identities in *Aux1* could be mapped to those in *NCHLT-clean*, initial data augmentation experiments could be conducted. The additional 27.8 hours of training data significantly improved Afr recognition results. In contrast, results seem to indicate that the 17.7 hours of additional Eng training data was not enough to achieve a similar improvement for Eng. These trends were successfully verified for both languages using data from different corpora.

Future work should focus on efficiently extending the training data sets for all languages. The *Aux1* and *Aux2* data contains many repetitions of the "search term-like" prompts in the *NCHLT-clean* train and test data sets. The impact of these repetitions on various neural models still needs to be assessed. The identity of the speakers in the *Aux2* data also has to be verified against the speakers in *NCHLT-clean* and *Aux1*.

VII. Acknowledgements

The research reported on in this paper was funded by the South African Centre for Digital Language Resources (SADiLaR). The authors are indebted to Mr Andrew Gill of the Centre for High Performance Computing for providing technical support.

References

[1] E. Barnard, M. H. Davel, C. van Heerden, F. de Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," in *Proccedings of the 4^{th} Workshop on Spoken Language Technologies for Under-resourced Languages*, St Peterburg, Russia, May 2014, pp. 194–200.

[2] M. H. Davel, C. van Heerden, and E. Barnard, "Validating smartphone-collected speech corpora," in *Proceedings of SLTU*, Cape Town, South Africa, May 2012, pp. 68–75.

[3] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK Book. revised for HTK version 3.4," March 2009, http://htk.eng.cam.ac.uk//.

[4] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on Automatic Speech Recognition and Understanding (ASRU)*, no. EPFL-CONF-192584, Hilton Waikoloa Village, Big Island, Hawaii, December 2011.

[5] J. Badenhorst and F. de Wet, "The limitations of data perturbation for ASR of learner data in under-resourced languages," in *Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), 2017*. IEEE, 2017, pp. 44–49.

[6] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proceedings of Interspeech*, 2015, pp. 3214–3218.

[7] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[8] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014.

[9] A. Biswas, F. de Wet, E. van der Westhuizen, E. Yılmaz, and T. Niesler, "Multilingual neural network acoustic modelling for ASR of under-resourced English-isiZulu code-switched speech," in *Proceedings of Interspeech*, 2018.

[10] Z. Yu, V. Ramanarayanan, D. Suendermann-Oeft, X. Wang, K. Zechner, L. Chen, J. Tao, A. Ivanou, and Y. Qian, "Using bidirectional LSTM recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 338–345.

[11] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and LSTMs," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 373–377, 2018.

[12] M. Karafiat, M. K. Baskar, K. Vesely, and J. C. F. Grezl, L. Burget, "Analysis of multilingual BLSTM acoustic model on low and high resource languages," in *Proceedings of IEEE ICASSP*, Galgary, Canada, April 2018, pp. 5789–5793.

[13] R. Eiselen and M. J. Puttkammer, "Developing text resources for ten South African languages." in *LREC*, 2014, pp. 3698–3703.

[14] N. J. de Vries, M. H. Davel, J. Badenhorst, W. D. Basson, F. de Wet, and E. B. an Alta de Waal, "A smartphone-based ASR data collection tool for under-resourced languages," *Speech Communication*, vol. 56, pp. 119–131, 2014.

[15] E. van der Westhuizen and T. R. Niesler, " Technical Report SU-EE-1501 An Analysis of the NCHLT Speech Corpora," Stellenbosh University of Zurich, Department of Electrical and Electronic Engineering, Tech. Rep., 05 2015.

[16] L. Loots, E. Barnard, and T. Niesler, "Comparing manually-developed and data-driven rules for p2p learning," in *Pattern Recognition Association of South Africa (PRASA)*. Stellenbosch, South Africa: IEEE, November 2009, pp. 35–40.

[17] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proceedings of Interspeech*, Dresden, Germany, September 2015, pp. 3586–3589.

[18] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalised maxout networks," in *Proceedings of IEEE ICASSP*, Florence, Italy, May 2014, pp. 215–219.

[19] D. Jurafsky and J. Martin, *Speech & language processing*. Prentice Hall, 2000.

[20] F. de Wet, A. de Waal, and G. B. van Huyssteen, "Developing a broadband automatic speech recognition system for Afrikaans," in *Proceedings of Interspeech*, Florence, Italy, August 2011, pp. 3185–3188.

[21] H. Kamper, F. de Wet, T. Hain, and T. Niesler, "Resource development and experiments in automatic South African broadcast news transcription," in *Proceedings of SLTU*, Cape Town, South Africa, May 2012, pp. 102–106.