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 CORPORAS

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. 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>
<thead>
<tr>
<th>Language</th>
<th>Aux1 init</th>
<th>Aux1 failed</th>
<th>Aux1 dur</th>
<th>Aux2 init</th>
<th>Aux2 failed</th>
<th>Aux2 dur</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afr</td>
<td>54 117</td>
<td>2 451</td>
<td>42.68</td>
<td>47 290</td>
<td>356</td>
<td>39.14</td>
</tr>
<tr>
<td>Eng</td>
<td>42 958</td>
<td>952</td>
<td>29.78</td>
<td>54 719</td>
<td>628</td>
<td>38.92</td>
</tr>
<tr>
<td>Nbl</td>
<td>37 669</td>
<td>3 224</td>
<td>42.56</td>
<td>100 402</td>
<td>4 202</td>
<td>120.07</td>
</tr>
<tr>
<td>Nso</td>
<td>65 224</td>
<td>2 259</td>
<td>64.89</td>
<td>53 318</td>
<td>947</td>
<td>51.80</td>
</tr>
<tr>
<td>Sot</td>
<td>74 457</td>
<td>5 858</td>
<td>73.86</td>
<td>47 938</td>
<td>700</td>
<td>43.51</td>
</tr>
<tr>
<td>Ssw</td>
<td>67 410</td>
<td>7 172</td>
<td>78.41</td>
<td>136 422</td>
<td>9 490</td>
<td>167.00</td>
</tr>
<tr>
<td>Tsn</td>
<td>69 655</td>
<td>1 953</td>
<td>70.15</td>
<td>35 156</td>
<td>356</td>
<td>36.98</td>
</tr>
<tr>
<td>Tso</td>
<td>71 311</td>
<td>3 781</td>
<td>83.67</td>
<td>2 316</td>
<td>1 489</td>
<td>0.65</td>
</tr>
<tr>
<td>Ven</td>
<td>82 895</td>
<td>4 886</td>
<td>93.69</td>
<td>44 666</td>
<td>1 220</td>
<td>54.94</td>
</tr>
<tr>
<td>Xho</td>
<td>90 560</td>
<td>8 739</td>
<td>102.95</td>
<td>53 269</td>
<td>2 549</td>
<td>54.95</td>
</tr>
<tr>
<td>Zul</td>
<td>77 833</td>
<td>3 471</td>
<td>97.93</td>
<td>30 319</td>
<td>327</td>
<td>32.74</td>
</tr>
<tr>
<td>Total</td>
<td>734 089</td>
<td>6.1%</td>
<td>780.57</td>
<td>605 815</td>
<td>3.7%</td>
<td>640.70</td>
</tr>
</tbody>
</table>

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 not in the training set are listed in the NCHLT TST Type column.

1Three 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), Setswana (Tsn), Xitsonga (Tso), Tshivenda (Ven), isiXhosa (Xho), isiZulu (Zul).
NCHLT TRN&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 Types 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-

²Type 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. 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.

![Graph of local phone error rates (PERs) for 400 utterance subsets of the Aux1 data.](image)

Figure 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 Afr and Eng systems). 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.

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%).

![Graph comparing PERs of all acoustic models on NCHLT test data.](image)

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


