Gauging the accuracy of automatic speech data harvesting in five under-resourced languages

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

de Wet, Febe
Department of Electrical and Electronic Engineering, Stellenbosch University and
School of Electrical, Electronic and Computer Engineering, North-West University, Potchefstroom, South Africa
fdw@sun.ac.za

Abstract

Recent research on deep-learning architectures has resulted in substantial improvements in automatic speech recognition accuracy. The leaps of progress made in well-resourced languages can be attributed to the fact that these architectures are able to effectively represent spoken language in all its diversity and complexity. However, developing advanced models of a language without appropriate corpora of speech and text data remains a challenge. For many under-resourced languages, including those spoken in South Africa, such resources simply do not exist. The aim of the work reported on in this paper is to address this situation by investigating the possibility to create diverse speech resources from unannotated broadcast data. The paper describes how existing speech and text resources were used to develop a semi-automatic data harvesting procedure for two genres of broadcast data, namely news bulletins and radio dramas. It was found that adapting acoustic models with less than 10 hours of manually annotated data from the same domain significantly reduced transcription error rates for speaking styles and acoustic conditions that are not represented in any of the existing speech corpora. Results also indicated that much more automatically transcribed adaptation data is required to achieve similar results.

Keywords: low-resource languages, automatic speech recognition, data harvesting, domain adaptation, data collection, broadcast data

1 Introduction

Speech is the most natural way for human beings to communicate with each other. It is often said that speech is what sets humans apart from other animals. It is therefore no surprise that being able to "talk over long distances" (the invention of the telephone) was regarded as a major technological breakthrough. Since the advent of the telephone and other communication devices, speech signals have been processed in many different ways. One of these is automatic speech recognition (ASR), also know as speech-to-text conversion. What started in the late 1970s in a few research labs around the world has grown into a multi-billion industry worldwide. If speech is what makes us human, then automatic speech processing is part of the digital manifestation of humanity. Moreover, having access to language in a digital format, either in its written or spoken form, has unlocked the possibility to study languages in many ways that were not possible before. The attempt to make more speech data available in South Africa's official languages described in this paper is therefore not only relevant for technology development, but also for studying the languages themselves.

Data harvesting is proposed as a solution to circumvent the data limitations that hamper ASR technology development in under-resourced environments. Our aim is to produce new speech corpora, including data in various acoustic environments and speaking styles. to unlock the full potential of state-of-the-art ASR in South Africa's languages. The data harvesting initiative is supported by the South African Centre for Digital Language Resources (SADiLaR[1]) as part of their efforts to develop indigenous language resources. A feasibility study focusing on the collection, building and

testing of automatic transcription systems yielded initial transcriptions of broadcast speech data in two languages (Badenhorst & de Wet 2021). The results of the study indicated that the proposed approach could also be used to collect data in other languages.

This paper elaborates on the feasibility study by determining how accurately broadcasts in four South African languages, Afrikaans (Afr), Tshivenda (Ven), isiZulu (Zul) and Sepedi (Nso) could be transcribed automatically. South African English (Eng) was also used in initial model adaptation experiments. In addition, the relative contribution of the text corpora that are available in each language is discussed.Although these resources are not extensive, attempts to utilise Kaldi RNN-based language modelling yielded transcription accuracy improvement.

Recent studies reported promising results when time-delayed neural network (TDNN) acoustic models were adapted using strategies such as transfer learning, re-training of entire networks for a few epochs and i-vector adaptation (Szaszák & Pierucci 2019). Similar techniques were used in this study to adapt existing speech systems based on factorised TDNN (TDNN-F) models (Badenhorst & de Wet 2019, 2022) with data from the harvesting domains. TDNN-F adaptation was initially tested by adjusting baseline English models with speech from the *government speeches* domain. The same technique was subsequently applied using adaptation data from other domains in Afr, Ven and Zul.

2 Data harvesting procedure

The data harvesting procedure we propose comprises four phases, each consisting of individual tasks. Figure 1 describes the procedure for a single language.

The availability of appropriate resources is crucial to automatic data harvesting. To produce the best possible annotations of collected data, appropriate techniques also need to be identified and their limitations considered. One such limitation is the mismatch between the acoustic properties of the data

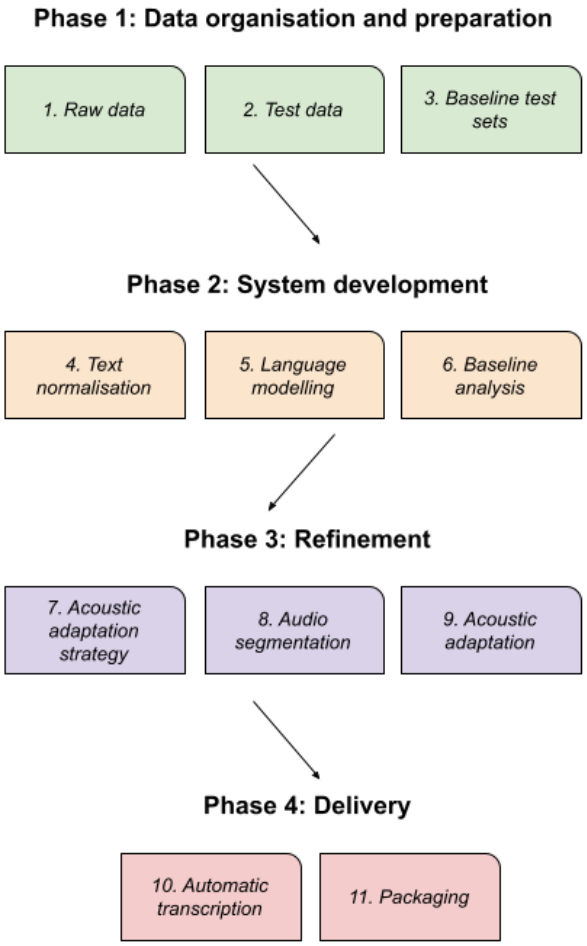

Figure 1: High level diagram of data harvesting procedure

on which the harvesting system was trained and those of the data in the target domain. Such mismatches invariably result in recognition errors and hence, in this particular application, incorrect transcriptions.

Phase 1 consists of three tasks related to data organisation and preparation. Task one focuses on acquiring raw data in such a way that initial models can be used but also keeping in mind that these models will require adaptation to other domains in subsequent phases of the procedure. Evaluating automatic annotation approaches requires test data sets from the target domain. Tasks two and three therefore focus on creating manually annotated test data sets. In the current investigation, test segments con-

taining music or overlapping speech from multiple speakers were excluded during evaluation.

Phase 2 of the data harvesting procedure focuses on ASR system development. Text normalisation prepares text data for model development. Design choices regarding the implementation of ASR vocabularies could, for example, require dates and numbers to be written out or the exclusion of email addresses etc. Processed texts enable various ASR modelling options, for example, word and sub-word transcription systems. Sub-word systems using speech phones as tokens were implemented in the investigation reported on here. The last step of Phase 2 involves configuring the automatic transcription system's initial acoustic and language models and analysing its performance.

The refinement performed in Phase 3 aims to adapt ASR acoustic models to the domain from which data is to be harvested. The acoustic differences between the training and target data should inform the choice of a suitable adaptation strategy. Speech recognition tools such as the Kaldi toolkit (Povey et al. 2011) rely on the ability to determine the time alignments of transcription labels given the audio. Transcription errors tend to increase the difficulty of finding these alignments, especially for longer pieces of audio. If at all tractable, the alignment process may become much slower. Fortunately, acoustic adaptation recipes can be designed that circumvent some of the above mentioned problems, such as using donor model alignments created for target data transcribed in Phase 2, and applying segmentation techniques to produce shorter segments of the target audio. In our experiments segmenting audio into relatively short segments enabled optimisation during the acoustic adaptation task. Confidence scoring was applied to adapt models using acoustically better matching segments first.

Phase 4 starts with an automatic transcription step, in which transcriptions are generated for all audio segments harvested from the target domain using appropriately adapted acoustic models. These transcriptions can be either sub-word or word representations. Both were used in this study. The data harvesting procedure is completed by packaging a subset of automatically transcribed audio together with relevant meta-data and documentation.

3 Data

For the purpose of speech technology development, broadcast speech data is a largely under-utilised source of data in South African languages. Initially, it was anticipated that broadcasts from various radio stations would be streamed to a server and recorded. This would be beneficial for resource development since the collection of raw data would scale to hundreds of hours of raw data which means that streaming could potentially provide *big data* over time.

However, during the process of identifying possible data sources, other options also emerged. For example, some radio stations offered to provide physical copies of their content while others were willing to transfer data via a dedicated link. Another possibility identified was to transfer data from a content hosting service in the form of podcasts. Different types of broadcast data could be obtained from each data capturing option. For example, some broadcasts contain news bulletins of high quality audio, but only for a limited number of news readers. News broadcast scripts were only available from some radio stations.

South African hosting services provides numerous radio show podcasts in all languages. An advantage of transferring data from a hosting service is that it enables podcast transfers in bulk. In Badenhorst & de Wet (2021) it was reported that, during the Covid-19 pandemic, collecting broadcasts from individual stations was severely impacted and getting data sharing agreements in place became an extremely slow and time-consuming process. In these circumstances collaboration with an online audio platform providing podcasts and audio live streaming services resulted in successful agreements with a number of radio stations. Particularly the drama podcasts from these stations could be utilised as test and adaptation data (during Phases 1 and 3 in Figure 1) from the conversational domain.

3.1 Audio data

3.1.1 Test data

The second task in the data harvesting procedure was to select and prepare test data. As the conclusions drawn from the Phase 2 baseline transcription would directly depend on each broadcast test data set, these sets needed to represent the domains from which data was going to be harvested. Considerations included the acoustic conditions and speaking styles represented in the raw data.

Two types of test data sets were created in each language. Studio news data was chosen as a better match to the baseline models because these models were trained on read speech segments. However, acoustic model development for automatic transcription would require training on more diverse data. The second test set for each language was therefore compiled from radio dramas, because they typically contain speech produced by multiple speakers in various acoustic conditions.

Complete news bulletins included news chimes and clips by reporters. Only the in-studio news reading segments were included in the test sets. Similarly, radio drama episodes contained a diverse mix of acoustic events such as music during introduction or transition sections. Background noise is also abundant and sometimes the conversation between characters overlap. Care was taken to include only clear speech parts without any overlap in the drama test data sets. All the test data was manually transcribed by a transcription company. In total, nine test data sets were created to enable the baseline analysis task (Phase 2 of the harvesting procedure). Table 1 presents the duration (in hours) as well as the speaker distribution for each test set.

The Eng Speeches test set comprises five-minute segments of audio, each originating from a speech delivered by the President of South Africa (Mr. Cyril Ramaphosa). The selected speeches were all delivered in the same year (2019-2020), but represent different event categories. Six of the twelve speeches were recorded indoors without an audience. These indoor speeches were similar to radio news bulletins with regard to speaking style and recording conditions. The other six speeches contain background noise due to the presence of an audience and breathing sounds from the speaker. The event categories of the Eng Speeches are as follows: the State Of the Nation Address (SONA), Lekgotla, Lunch, Commemoration, National Day as well as Official Opening. The National Day was an outdoor event and the Official Opening was recorded with a bad microphone setup. The Phase 3 adaptation strategy design task was first tested on this data to validate the adaptation process before proceeding with acoustic model adaptation in the other languages.

Table 1: Duration and speaker distribution of test data

Test set	Duration (h)	Speaker distribution
Eng Speeches	1.00	1 male
Afr News	7.89	18 male, 10 female
Afr Messages	0.36	4 male, 4 female
Afr Drama	0.82	multiple
Ven News	0.54	3 male, 2 female
Ven Drama	0.54	multiple
Nso News	0.46	1 male, 3+ female
Nso Drama	0.60	multiple
Zul News	0.51	3+ male, 4 female
Zul Drama	0.67	multiple

A relatively large Afr News test set could be selected from an existing Afr corpus (De Wet et al. 2011). In addition, the Afr Messages data, introduced in Badenhorst & de Wet (2021), was included as an example of Afr studio speeches.

3.1.2 Adaptation data

Four sets of adaptation data were used to reduce the mismatch between data from the target domains and the transcription systems' acoustic models. An initial investigation utilised a subset of the NCHLT II Eng parliamentary speech corpus (De Wet et al. 2016). The NCHLT II data set's duration of 50 hours is similar to that of the NCHLT I Eng cor-

pus. The segmentation and ranking that was used to select the adaptation data is discussed in Section 4.3. Phone-based dynamic programming (PDP) ranking of the NCHLT II segments relative to the NCHLT I models was performed before adaptation. Utterances with better matching acoustics and transcriptions were prioritised during adaptation.

In addition to the Afr Messages test set, another 85 hours of automatically transcribed Afr Messages data was available. It was deemed a good data set for acoustic model adaption since the studio quality podcasts only contained the speech of a single speaker per episode and no other acoustic events such as start and end chimes, introductions or any music clips.

Two more adaptation data sets were compiled from Ven news and Zul drama data respectively. News data was chosen as an example of data that matches the NCHLT acoustics on which the baseline models were trained relatively well. Approximately 90 hours of Ven news bulletins were available before segmentation and PDP ranking. Lastly, voice activity detection (VAD) segmentation was applied to approximately 25 hours of Zul drama episodes extracting eight hours of speech segments. Each segment was manually inspected by a transcription company, after which approximately six hours of data were manually transcribed based on annotation markers. At the time of writing no adaptation data was available for Nso.

3.2 Text data

Data harvesting requires text data that adequately covers the vocabulary of the speech in the target domain. Modelling the vocabulary from available texts not only enables word transcription, but also sub-word phone transcription systems. In well-resourced languages texts containing millions and for some languages even billions of words are incorporated in language models. The current investigation used all the text available in the four target languages. The corpora were developed during the Lwazi (Calteaux et al. 2013), NCHLT Speech[2], NCHLT Text (Eiselen & Puttkammer 2014) and Autshumato (McKellar & Puttkammer 2020, Groenewald & du Plooy 2010) projects. From the Lwazi projects, the TTS text corpora were chosen because they include phonetically-balanced sentences that contain very few out-of-language tokens.

Table 2 compares the sizes of the text corpora in the four target languages. The name of each corpus appears in the first row of the table. The number of unique words (N=1) gives an indication of the vocabulary size while the word 2-gram counts (N=2) correspond to word sequences. The total number of words for each corpus (T) denotes the corpus size. As is indicated in the table, the NCHLT Text corpora are much larger than the Lwazi TTS and NCHLT Speech data sets. For the Autshumato texts only the Parallel Text corpora are comparable in size to the NCHLT Text corpora. Apart from these, the Autshumato machine translation (MT) evaluation text (Eval in Table 2) and multilingual word and phrase lists were used to add additional words to the transcription system's vocabulary.

The values in Table 2 indicate that the Afr Lwazi TTS text corpus contains more unique words (12 447) than the annotations of the NCHLT Speech data. The table also shows that the vocabulary size for the Afr NCHLT text corpus is almost four times that of the Lwazi TTS text corpus. Although the Afr component of the Autshumato Parallel text corpus consists of a similar number of words than the NCHLT text corpus, its vocabulary size is substantially smaller (30 440 unique words).

In contrast, the Ven Lwazi TTS text corpus contains only 3 488 unique words, fewer words than the Ven NCHLT Speech data annotations. While the NCHLT Text corpus contains more than seven times the vocabulary of the Lwazi TTS text corpus, compared to the Afr component, the vocabulary size is about half the number of unique words in Afr. The Nso Lwazi TTS text corpus contains 5 125 words, more than Ven, but less than half the number of Afr words. However, the Nso NCHLT

Table 2: Comparison of vocabulary sizes for different text data sources

		Lwazi	NCHLT		Autshumato			
Language	**N**	**TTS**	**Speech**	**Text**	**Words**	**Phrases**	**Eval**	**Parallel**
Afr	1	12 447	8565	56 192	11 785	1226	3015	30 440
	2	66 652	18 205	424 438	61	1061	11 126	167 830
	T	**143 958**	**173 128**	**2357 560**	**11 892**	**2508**	**38 125**	**2341 627**
Ven	1	3488	7578	24 314	4435	994	2877	-
	2	12 134	26 135	198 136	62	1636	15 132	-
	T	**30 835**	**217 526**	**996 393**	**4899**	**3894**	**45 513**	-
Nso	1	5125	11 055	57 400	4938	1006	3542	26 540
	2	21 602	25 124	384 963	90	1681	15 598	148 526
	T	**50 952**	**266 261**	**2209 452**	**5190**	**4160**	**47 250**	**843 017**
Zul	1	12 367	23 911	185 290	6358	1621	9074	72 145
	2	23 147	17 304	900 785	142	1606	19 339	233 986
	T	**27 288**	**114 050**	**1555 103**	**6725**	**2931**	**28 909**	**413 050**

Text corpus is larger and compares well with the Afr NCHLT Text corpus size. Although almost all the Zul texts contain the largest vocabulary for the particular type of text, this trend does not correspond to high T values. This discrepancy can be ascribed to the fact that Zul is an agglutinative language.

4 Harvesting system configuration

The configuration of the transcription system for Phase 2 (System development) and Phase 3 (Refinement) of the harvesting procedure is described in the following sections. The description includes the choices made with regard to text normalisation, subsequent language modelling refinements as well as segmentation and adaptation strategies that can be applied without the initial support of a language model.

4.1 Text normalisation

A generic text normalisation procedure was followed to convert each of the text resources to the required representations for further processing. Preprocessing was applied to remove empty lines, lines containing document headers, bulleted lists etc. For example, some texts employed capitalisation to indicate spelled-out words and names. Best effort manual verification of these tokens was performed. A decision was made to split all detected spelled-out tokens into single characters. This procedure ensured that the text would generate fewer pronunciation errors as it simplifies the pronunciation dictionary extension process significantly. All characters were converted to lowercase.

To clean the text further, the text normalisation protocol of the locally developed Speect TTS software was applied (Louw et al. 2010). This software applies a rule-based, number spell-out algorithm (Gillam 1998) that, for example, writes out numbers, monetary amounts and time in each language. The Speect normalisation engine (Louw et al. 2016), also consistently removed many unwanted lines (containing URLs and other graphemes such as strings of dashes etc.). Subsequently, character checks were performed against the list of allowed graphemes in the grapheme to phoneme (G2P) rule set for each language (Barnard et al. 2014). Any remaining incompatible graphemes in the text were either replaced or removed semi-automatically.

4.2 Language modelling

Section 3.2 introduced and compared text resources in the four South African languages included in this study. These limited resources are approximately 1 to 3 million words in size, which is a limitation for word transcription system development. The following sections describe how these resources were used in combination with standard n-gram language modelling (Goodman 2001) to implement speech data harvesting. Sections 4.2.1 and 4.2.2 describe the development of n-gram models for subword and word level text representations while Section 4.2.3 describes a Kaldi RNNLM setup applying TDNN-LSTM language models for word recognition.

4.2.1 Phone n-gram models

To produce accurate phone transcriptions, the harvest setup employed transcription systems with phone language models built from combinations of the texts in Table 2. To determine whether combining the available texts would result in better coverage of phone contexts we computed n-gram phone label counts for the text resources. Table 3 provides counts of the additional unique biphone labels a second set of text contributes if the Lwazi TTS text in each language is used as the point of departure. The n-gram counts for the Lwazi TTS text are provided in the second column of Table 3. The TTS texts consist of phonetically balanced sentences, so they should provide good phone coverage for all four languages. Preliminary tests confirmed good phone transcription results using these texts only. Subsequent columns in Table 3 report the additional unique labels available from the other texts. For the NCHLT texts, one column also reports three texts combined together, showing the contribution that the NCHLT Parallel Texts provide with the NCHLT Text Corpus already included.

The values in Table 3 show that combining the Lwazi TTS texts with the NCHLT Text corpora leads to the largest contributions of additional phone contexts. This seems to be especially true for the Ven, Nso and Zul languages, where the number of unique biphone contexts are effectively doubled combining the resources. Interestingly, a similar result is seen for Nso and Zul if the Autshumato Parallel texts are used instead. Combining the larger Afr TTS texts with the Autshumato Parallel text provides a much smaller contribution (only 174 additional biphone labels). The values in Table 3 show that, if the NCHLT Text corpora are already included (TTS + Text + Parallel), further contributions from the Autshumato Parallel text to the counts are limited. Similarly, contributions from other Autshumato texts given the TTS texts are smaller, but do include some additional contexts.

4.2.2 Word n-gram models

Another indication on the sufficiency of the text resources for data harvesting development is the out-of-vocabulary (OOV) rates given representative test data from the harvest domain. We analysed the OOV for all harvest languages. Figure 2 presents OOV rates for Afr, Ven and Nso as the percentage of words in the test data sets that are not in the vocabulary of four different texts: the Lwazi TTS text (TTS), NCHLT Text corpora (Text), a combination between the Lwazi TTS and the NCHLT Text corpora (TTS+Text) and lastly a greedy selection (*Greedy*) given all of the available text resources. The *Greedy* data set was compiled using an algorithm that selects lines of text based on whether a line includes new vocabulary words or not. If a line only contains already selected words, it is discarded. Text was selected in this manner to limit the skewing of word context distributions that could be caused by simply pooling all the text. Starting with the phonetically balanced TTS Texts, the remaining texts were processed in the following order: the Autshumato text resources, the NCHLT text, NCHLT Speech and Autshumato Parallel texts.

Figure 2 clearly shows that only utilising the Lwazi TTS text leads to the highest OOV rates for the different test data sets (between 10% and 20% OOV). Including vocabulary from the NCHLT texts is effective since OOV rates are substantially reduced.

Table 3: Bi-gram phone contribution of unseen unique phone labels for different text combinations

Language	NCHLT					Autshumato		
	TTS	TTS + Speech	TTS + Text	TTS + Text + Parallel	TTS + Words	TTS + Phrases	TTS + Eval	TTS + Parallel
Afr	1064	46	209	17	20	4	13	174
Ven	492	240	590	-	21	53	96	-
Nso	550	176	646	21	33	29	89	533
Zul	682	202	710	10	71	68	91	560

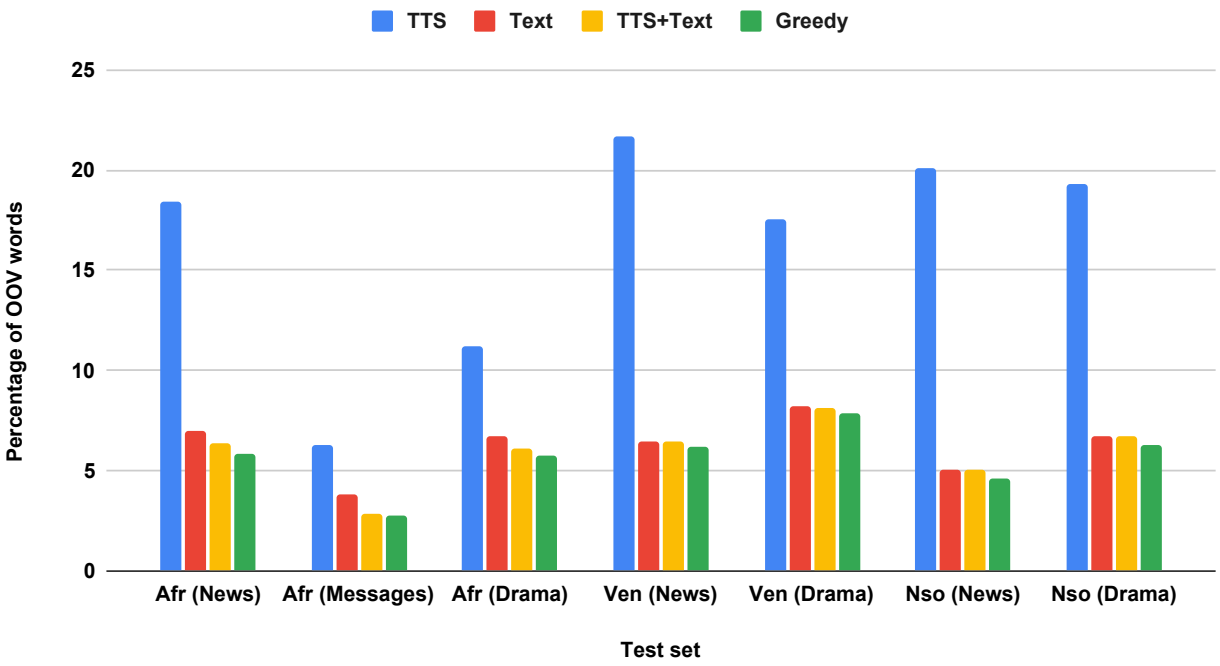

Figure 2: Out-of-vocabulary rates (as a percentage) of ARPA language models built from different texts and given the test data of the three languages.

Pooling the Lwazi TTS texts with the NCHLT Text corpora texts sometimes leads to slightly lower OOV rates. As expected, in all cases, the greedy selection provides the lowest OOV, usually about half a percent lower than the second lowest value.

Predicting the transcriptions of the test data employing trigram ARPA language models, Figure 3 provides the corresponding perplexity estimates for models compiled with each set of text data. The figure shows higher perplexities for most of the Lwazi-TTS-text-only models. It can also be seen that n-gram models based on the NCHLT Text corpora achieve fairly good perplexity values. Models based on a combination of the Lwazi TTS and the NCHLT Text corpora (TTS+Text) as well as the greedy selection (Greedy) strategies do not lower perplexity values much further, but yield similar perplexity. The Afr Messages and Afr Drama test set transcriptions are exceptions to this rule. Here the models derived from the larger (compared to other languages, see Table 2) Lwazi TTS text generate better perplexity estimates than models based

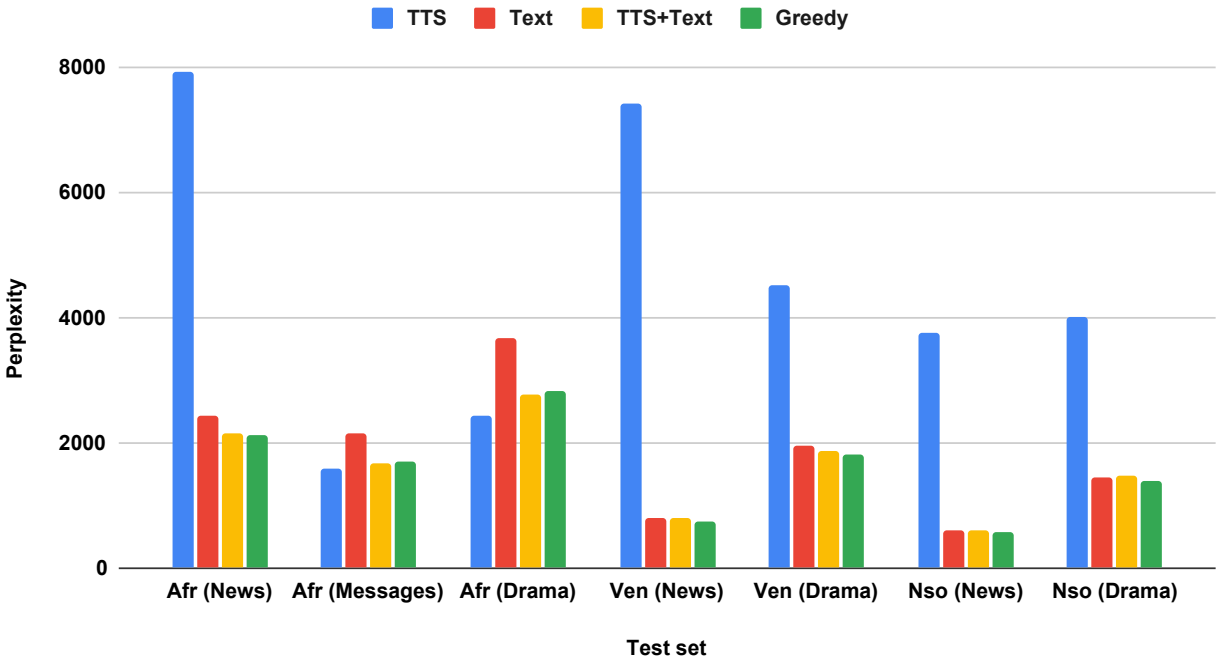

Figure 3: Perplexity estimates of 3-gram language models built from different texts and given the test data of the three languages.

Table 4: OOV word rates and perplexity estimates for tri-gram language models derived from the Zul texts

	Lwazi	NCHLT		Other
Test set	**TTS**	**Text**	**TTS + Text**	**Greedy**
		OOV word rates		
Zul News	49.09%	16.26%	16.26%	15.87%
Zul Drama	44.58%	17.13%	17.05%	16.56%
		Perplexity estimates		
Zul News	551 182	47 627	47 839	46 768
Zul Drama	334 283	63 385	63 425	63 260

on the Afr NCHLT Text corpora. Lastly, given the larger NCHLT Text data, the Drama test transcriptions seem to generate larger perplexities than News data for all languages, alluding to different vocabularies in the transcriptions of these conversational test data sets.

With similar sized texts, the larger Zul vocabulary results in larger OOV word rates. In fact, OOV word rates for Zul given the same text resource types are more than twice as high as for the other languages. Perplexity estimates for Zul are also much higher than for the other languages. Given these differences the Zul OOV and perplexity estimates are reported separately in Table 4. Similar trends are observed with the NCHLT Text improving values considerably. Again, for the Drama test transcriptions, higher perplexity values compared to the News test transcriptions are observed for models including more text.

4.2.3 Kaldi RNNLM

Kaldi-RNNLM is an extension of Kaldi that supports neural network language models (Xu, Li, Wang, Wang, Kang, Chen, Povey & Khudanpur 2018). The recipe incorporates sub-word modelling using letter n-gram based features, improving the modelling of rarely seen word contexts. Subsequent ASR systems (Chen et al. 2018) also employ LSTM-based language models (Xu, Li, Wang, Wang, Kang, Chen, Povey & Khudanpur 2018, Jozefowicz et al. 2016). These models are even more effective at

modelling wide contexts than conventional RNNs. Kaldi uses a two-pass decoding method to apply the models, since RNNLMs encode a theoretically infinite history length (Xu, Chen, Gao, Wang, Li, Goel, Carmiel, Povey & Khudanpur 2018). The recipe in Xu, Li, Wang, Wang, Kang, Chen, Povey & Khudanpur (2018) employs an n-gram approximation lattice-rescoring method. First a representative n-gram model is used to compile and decode, resulting in a decoding graph. This generates a set of possible hypotheses, after which lattice-rescoring can be applied.

One version of the harvesting system includes an RNNLM recipe which is based on the Kaldi Switchboard example. The recipe demonstrates a forward (Povey 2018*b*) and backward (Povey 2018*a*) TDNN-LSTM. The backward model training requires reversed text at the sentence level (Arora et al. 2020). To implement the backward model, the forward model first rescores the original trigram decode lattices that correspond to the NCHLT Text based ARPA model (see Section 4.2.2). Subsequently, another rescoring of the backward model on top of the already rescored forward model lattices is performed. Adjusting the interpolation weight optimises the recognition results for the rescoring steps.

As in Yang et al. (2018) the chosen recipe trains 5-layer TDNN-LSTM LMs for rescoring, alternating the TDNN and LSTM layers. Each TDNN layer combines the current and previous time steps. The LSTM layers have hidden projection layer dimensions of 265 and the TDNN embedding dimensions remain set at the value of 1 024. In both the Nso and Zul languages, lines of the NCHLT Text Corpus text are selected as sentences utilising one line out of every 50 lines (approximately 2%) as development data. Keeping the maximum n-gram order parameter at the standard setting, implementing a 4-gram approximation allows 3-word histories to be kept intact.

4.3 Segmentation

The segmentation of acquired raw audio data and the subsequent segment selection of data suitable for automatic harvesting is necessary to support model adaptation. Different segmentation techniques were applied depending on the data set. In Badenhorst & de Wet (2021) the segmentation and selection of read speech resources such as radio news bulletins was performed in a two step process. Step one applied speaker diarisation in an unsupervised manner combined with a heuristic selection of start segments. This enabled detection of 90% of the news start times.

News bulletins included news clips, mostly in English or another language and usually the news ended with news chimes and music. To select the within language speech from this data, a second Kaldi alignment-based silence detection method was implemented. As mentioned in Section 2, it was possible to find label alignments for relatively long pieces of audio using a well estimated acoustic model. The Kaldi alignment also inserted silence phones into the forced alignment phone string automatically when required. Subsequently, segmentation was possible for silence labels of 0.1 seconds or longer in duration. Segmenting on selected silence labels, short segments of audio between 5 and 15 seconds in duration were produced. Optimised model estimation can be achieved by training on segments with good acoustic match first. Similar to the acoustic selection in Badenhorst & de Wet (2019), PDP scoring can be applied to rank segments. The cross-language news clips and music segments ranked lower than clear in-language speech segments.

Segmentation of drama data was also required. Broadcast drama episodes generally contain long pieces of music transitioning between scenes and various background effects. A VAD technique to determine where clear speech occurred seemed like a more appropriate technique to segment this data. The hybrid convolutional neural network-bidirectional long short-term memory network (CNN-BiLSTM) structure propsed in Wilkinson

& Niesler (2021) was used for this purpose. After setting a speech versus non-speech threshold of 0.7, sufficient speech segments between 5 and 15 seconds in length could be selected for model adaptation experiments. Adjacent short voiced segments were merged if the non-speech duration between them were 2 seconds or shorter and the merged segment had a duration of less than 15 seconds.

4.4 Adaptation

Kaldi allows for separate acoustic and language model development. This means the data harvesting process can generate sub-word transcriptions to enable acoustic model adaptation for short segments of speech data using a relatively weak language model. Therefore, short text prompts such as those corresponding to manual transcriptions of in-domain speech data or TTS text, can be used to adapt transcription systems.

Acoustic model adaptation within the Kaldi nnet3 framework can be accomplished in a number of ways. The effectiveness of any adaptation strategy relies on the acoustic properties of the adaptation data compared to that of previous training data. In this study different combinations of re-training and i-vector adaptation were evaluated. Not all of the adaptive training options could successfully be applied given the particular combination of donor model and harvest data. The diagram in Figure 4 illustrates the adaptive training configuration options. The best results were obtained when the initial NCHLT models were re-trained but the NCHLT i-vectors were kept intact.

Kaldi nnet3 acoustic modelling recipes rely on the ability to determine time alignments of transcription labels given the audio before training. The required input alignments are usually obtained from triphone Hidden Markov Model (HMM) systems. In this study the Kaldi HMM systems were derived from the same NCHLT training setup that was used in Badenhorst & de Wet (2019). The triphone alignments block in Figure 4 refers to triphone alignments produced by these systems.

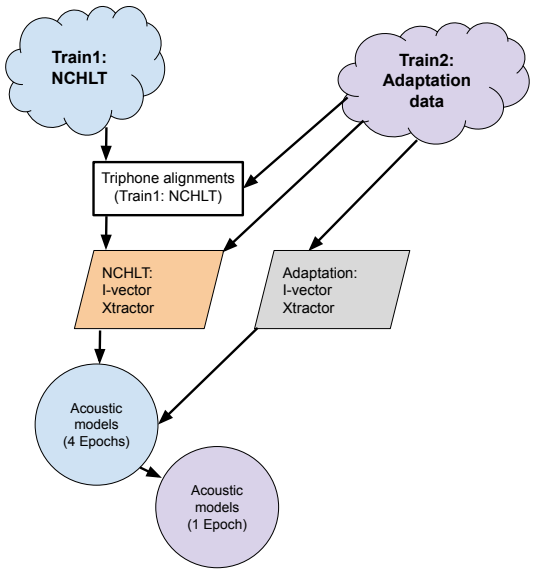

Figure 4: TDNN-F adaptive training configuration options. Different i-vector extractors can be applied to acoustically normalise the adaptation training data before training an additional epoch to adapt the NCHLT acoustic models.

The TDNN-F model adaptation recipe consisted of two stages. The first stage of the recipe reproduced the training setup of the initial model, but instead of finalising the process, the training setup remained in such a condition that a second stage of training could be applied (blue circle in the figure). The second stage is an adaptation stage, re-training the standard (four-epoch) model for an additional epoch using the adaptation data.

To enable standard feature extraction and model training, triphone alignments and data perturbation were applied to both the NCHLT and adaptation data. The initial NCHLT models were used to perform triphone alignment. The best adaptation results were obtained using the NCHLT i-vector extractor in both stages, generating i-vectors for the adaptation data with this extractor. It was also advantageous since no speaker information for the broadcast drama or NCHLT II adaptation data (for which the segment speakers were unknown) was available during adaptation.

Re-training required control of the intensity of the

adaptive training. This can be achieved by adjusting the learning rate for the training iterations of the last epoch, the number of iterations or training for more epochs. To obtain the results reported in this study, similar thresholds to the lower thresholds of training (0.000020 and 0.000015 respectively) were set during the last epoch. This restricted the algorithm to apply a learning rate close to the lower setting for the remaining iterations of training.

5 Results

To evaluate the performance of various harvest system configurations, transcription of the test data sets described in Section 3.1.1 were performed. Initially, sub-word harvest systems were built (Section 5.1) employing phone language models derived from combinations of the texts described in Section 3.2. This was followed by building the word harvest systems described in Section 5.2. Most sub-word harvest systems could later be refined further given the word harvest systems (Section 5.3). Section 5.4 describes a number of acoustic adaptation experiments.

5.1 Baseline sub-word harvest systems

The phone error rate (PER) values in Table 5 compare three harvest systems employing: 1) flat phone-based ARPA language models, 2) 3-gram phone-based ARPA language models and 3) 6-gram phone-based ARPA language models. As explained in Section 4.2.1, the 3-gram and 6-gram models were derived from the combination of texts that yielded the best performance. Both the 3-gram and 6-gram ARPA phone models were built with TTS text for Afr, TTS and NCHLT Text for Ven and NCHLT texts only for Nso and Zul. NCHLT acoustic models were used in all the baseline harvesting systems.

Table 5: Baseline sub-word harvest system results: Phone error rates for Flat, 3-gram and 6-gram phone language models.

	ARPA		
Test set	**Flat**	**3-gram**	**6-gram**
Afr News	20.98	18.76	16.34
Ven News	27.12	24.45	20.84
Nso News	29.50	27.27	24.05
Zul News	38.81	33.10	30.55
Afr Messages	25.82	24.15	22.42
Afr Drama	42.69	40.86	39.52
Ven Drama	45.80	42.40	40.35
Nso Drama	43.53	40.75	38.98
Zul Drama	46.66	42.96	41.58
Zul Drama Adapt	48.56	-	-

The values in Table 5 show that, in general, the harvest systems were much better at transcribing news data than drama data. The results also show that 3-gram systems outperform flat model systems in all cases and this trend continued for all 6-gram systems. The choice to use a 6-gram context size was made in Badenhorst & de Wet (2021) because lower PER rates could be achieved for TTS and larger sets of text data. The results indicate that transcribing the Afr Messages data resulted in more recognition errors than Afr News, but it was still much better than transcribing Afr Drama. Automatically transcribing the Zul adaptation data (Zul Drama Adapt) with a flat phone model yielded a comparable result to the Zul Drama test data, verifying the set's manual transcription accuracy.

5.2 Word harvest systems

Considering the various word level n-gram options reported in Section 4.2.2 for the text data introduced in Section 3.2, the word error rates (WERs) reported in Table 6 were obtained employing 3-gram ARPA language models. The values in the table show that including the NCHLT Text data when transcribing news data made a significant difference. This distinction was less pronounced for

the conversational messages and drama test data. It was also clear that including additional vocabulary with greedy selection resulting in smaller training texts did not degrade results. The worst word transcription rates were obtained for Zul.

Table 6: Baseline word harvest system results: Word error rates with 3-gram language models derived from different texts.

Test set	Lwazi		NCHLT		Other
	TTS	Speech	Text	TTS Text	Greedy
Afr News	48.72	62.27	37.42	36.25	**35.97**
Ven News	58.67	52.27	**42.22**	42.23	42.35
Nso News	60.54	63.31	47.78	47.67	**47.59**
Zul News	82.16	84.80	**69.24**	69.45	69.32
Afr Messages	47.08	61.87	46.01	**43.86**	43.94
Afr Drama	**67.63**	78.48	69.99	68.21	68.56
Ven Drama	75.82	78.39	72.16	71.73	**71.50**
Nso Drama	73.32	75.50	**71.23**	71.32	72.14
Zul Drama	86.68	90.95	83.42	83.40	**83.36**

An updated set of results were obtained by rescoring the Nso and Zul word trigram decode lattices with TDNN-LSTM language models. Table 7 compares the resulting ARPA and Kaldi RNNLM WERs for TDNN-LSTM language models developed only on the NCHLT text data sets. By adjusting the interpolation weight small but consistent improvements could be obtained for both the news and drama test data sets.

Table 7: Word harvest system results: Comparing 3-gram ARPA and Kaldi RNNLM WERs with the NCHLT Text Corpora in two languages.

Test set	ARPA 3-gram	RNNLM	Interpolation weight
Nso News	47.78	47.36	0.1
Zul News	69.24	68.36	0.2
Nso Drama	71.23	68.53	0.3
Zul Drama	83.42	81.82	0.3

Table 8: Refined sub-word harvest system results: Comparing PERs of 6-gram phone ARPA, 3-gram word ARPA and word Kaldi RNNLM systems.

Test set	Phone level	Word level	
	ARPA 6-gram	ARPA 3-gram	RNNLM
Afr News	16.34	**13.03**	-
Ven News	20.84	**18.09**	-
Nso News	24.05	21.26	**20.64**
Zul News	30.55	29.03	**28.12**
Afr Messages	22.42	**21.09**	-
Afr Drama	39.52	**38.56**	-
Ven Drama	**40.35**	41.28	-
Nso Drama	38.98	40.09	**37.91**
Zul Drama	**41.58**	47.00	43.42

5.3 Refined sub-word harvest systems

Interestingly, even lower PERs could be obtained for most test sets by first performing a word transcription. To convert the word transcriptions to phone labels a pronunciation dictionary was used. With the exception of the Ven and Zul Drama, the refined PER values reported in Table 8 indicate significantly lower error rates. Furthermore, Nso and Zul News as well as Nso Drama results benefited additionally with the Kaldi RNNLM transcription.

5.4 Acoustic adaptation

The adaptive training configuration defined in Section 4.4 was first tested on the Eng Speeches test set. Subset selections of the adaptation data were made based on per segment PDP scores. Table 9 shows that the PERs measured for the adapted acoustic models are significantly lower than those achieved by the baseline system (0% Adaptation data). Only 5 hours of adaptation data (10% Adaptation data) lowered the PER by more than 7%. The best adaptation result, a PER of 27.82%, was obtained when 60% of the NCHLT II data was used to adapt the acoustic models trained on NCHLT

Table 9: Phone error rates of Baseline NCHLT I English (0%) and adapted NCHLT I acoustic models using various percentages of the 50.19 hours of NCHLT II data, measured on the Eng Speeches test set.

% Adaptation data	ARPA Flat
0	49.33
10	41.72
20	36.91
30	31.62
40	29.90
50	28.26
60	27.82
70	31.08

Table 10: Adaptation results for six different types of speeches made by the President.

Event	Adapt 0%	Adapt 60%
SONA	45.70	18.05
Lekgotla	44.45	18.59
Launch	51.67	20.76
Commemoration	57.13	21.08
National day	69.61	36.58
Official opening	86.41	85.74

I data. This value compares well with the baseline News PERs reported for the four other languages in Table 5.

To determine whether acoustic adaptation was more effective for some speeches than others, PERs were calculated for six different types of data in the Eng speeches test set. The results are shown in Table 10. The most significant reduction in PER was obtained for the parliamentary data (SONA) and the indoor speeches (Lunch and Commemoration). This observation is expected, given that the NCHLT II data is similar to the data in these categories. Acoustic mismatch did, however, still restrict the impact of adaptation as is evident from the outdoor National Day and Official Opening results.

Table 11 provides an overview of the adaptation results applying the Afr, Ven and Zul adaptation data sets described in Section 3.1.2. Using PDP, the 10% and 20% best ranking segments of the Afr Messages adaptation data were selected corresponding to 8.69 and 17.39 hours respectively. Similarly, selecting 6% and 12% of the best matching Ven News adaptation data amounted to 7.83 and 15.66 hours. The phone transcriptions required for adapting the Afr and Ven models were derived using the baseline 6-gram phone transcription systems described in Section 5.1. The 6.12 hours manually transcribed Zul Drama adaptation data was also employed to train a new Zul acoustic model. Both the flat and 6-gram phone language model results for the adapted systems were compared to the baseline results presented in Table 5.

Table 11: PERs on the broadcast test data sets for Afr, Ven and Zul after acoustic adaptation.

Test set	Adapt data	ARPA Flat	ARPA 6-gram
Afr News	Messages 8.69h	20.70	17.51
Afr News	17.39h	20.94	17.43
Afr Messages	8.69h	20.50	17.81
Afr Messages	17.39h	20.20	17.37
Afr Drama	8.69h	43.82	41.74
Afr Drama	17.39h	43.95	41.24
Ven News	News 7.83h	22.25	19.51
Ven News	15.66h	22.37	19.63
Ven Drama	7.83h	46.54	43.72
Ven Drama	15.66h	46.34	42.84
Zul News	Drama 6.12h	32.68	**27.94**
Zul Drama	6.12h	30.33	**26.76**

Adapting with just 10% of the available data significantly improved the Afr Messages test set results for both Flat and 6-gram systems. However, doubling the adaptation data did not result in much further improvement. A similar observation was made for the Ven News adaptation with test data from the same domain. In these tests, adaptation based on the automatically transcribed data only yielded a marginal improvement for one test data set from the other domains, such as radio drama.

The Zul results in Table 11 clearly show that using manually transcribed adaptation data had the

most substantial impact on the transcription error rate. An absolute improvement of more than 14% PER was achieved transcribing the Zul Drama test data. This translates to an equivalent WER value of 65.45% (an absolute reduction of 18% compared to the corresponding value in Table 6) using the NCHLT Text data ARPA language model. Moreover, significantly lower PERs for the News data were also measured and a corresponding WER of 65.48% was obtained using the same language model.

6 Conclusion

In Badenhorst & de Wet (2021) it was observed that acoustic match to the harvest data is an important factor for transcription accuracy. The four baseline sub-word systems that were evaluated in this study confirmed this observation. NCHLT ASR training data only includes short read speech prompts. In comparison, the news test data contained much longer utterances and the conversational speech found in radio dramas included noisy acoustic environments. As expected, transcribing Drama data with NCHLT acoustic models produced larger transcription errors than News data. Word transcription results followed the same trend.

The results reported in this paper demonstrates the value of limited text resources to build sub-word transcription systems that can create a first transcription of unannotated broadcast data. Subsequent acoustic adaptation employing these transcriptions succeeded in lowering automatic transcription error rates. For the available text data, including vocabulary from the larger NCHLT texts provided the biggest benefit by lowering OOV rates considerably. However, high perplexity and OOV rates as well as high WERs confirmed the inadequacy of the Zul text to represent the language's larger agglutinative vocabulary. Novel solutions need to be sought to lower transcription error for agglutinative languages.

Further transcription accuracy improvement was achieved by applying TDNN-LSTM language models using the NCHLT text data to build an RNNLM. Sub-word transcription error could also be lowered further by converting the word transcription back to phones using a dictionary lookup. Re-evaluating the converted sub-word transcriptions also applying an adapted acoustic model would presumably lead to even larger error reductions.

The TDNN-F adaptation strategy proved to be effective. Two approaches to transcribe adaptation data were tested and the value of both was confirmed for different domains of speech data. Firstly, acoustic adaptation for the Ven News and the Afr Messages data was possible using automatic sub-word transcription only, proving that acoustic development can start with less than a hundred hours of raw unannotated data. Furthermore, the analysis showed how this process can be accelerated by manually transcribing a limited set of five or more hours of adaptation data. Even though the transcribed drama audio contained background noise and rapid conversational speech, acoustic adaptation could be performed effectively. Moreover, manually transcribing more spontaneous speech as adaptation data seems to enable cross-domain adaptation, because the adapted models also yielded improved transcription accuracy for the news test data. Future work should focus on refining cross-domain adaptation strategies and investigate the interplay between different speaking styles and limited quantities of training and adaptation data.

Notes

[1] `https://www.sadilar.org/index.php/en/`

[2] NCHLT Speech text comprises the ASR training prompts.

Acknowledgements

The data harvesting work 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

Arora, A., Raj, D., Subramanian, A. S., Li, K., Ben-Yair, B., Maciejewski, M., Zelasko, P., Garcia, P., Watanabe, S. & Khudanpur, S. (2020), The JHU Multi-Microphone Multi-Speaker ASR System for the CHiME-6 Challenge, *in* 'Proc. 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)', pp. 48–54.
URL: *http://www.isca-speech.org/archive/chime_2020/arora20_chime.html*

Badenhorst, J. & de Wet, F. (2019), 'The usefulness of imperfect speech data for ASR development in low-resource languages', *Information* **10**(9).
URL: *https://www.mdpi.com/2078-2489/10/9/268*

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).
URL: *https://upjournals.up.ac.za/index.php/dhasa/article/view/3820*

Badenhorst, J. & de Wet, F. (2022), 'NCHLT Auxiliary speech data for ASR technology development in South Africa', *Data in Brief* p. 107860.
URL: *http://www.sciencedirect.com/science/article/pii/S2352340922000725*

Barnard, E., Davel, M. H., van Heerden, C., de Wet, F. & Badenhorst, J. (2014), The NCHLT speech corpus of the South African languages, *in* 'Proceedings of the 4^{th} Workshop on Spoken Language Technologies for Under-resourced Languages', St Petersburg, Russia, pp. 194–200.
URL: *http://repository.nwu.ac.za/handle/10394/26493*

Calteaux, K., De Wet, F., Moors, C., Van Niekerk, D., McAlister, B., Grover, A., Reid, T., Davel, M., Barnard, E. & Van Heerden, C. (2013), Lwazi ii final report: Increasing the impact of speech technologies in south africa., Technical report, CSIR, Pretoria.
URL: *https://researchspace.csir.co.za/dspace/handle/10204/7138*

Chen, S.-J., Subramanian, A. S., Xu, H. & Watanabe, S. (2018), Building State-of-the-art Distant Speech Recognition Using the CHiME-4 Challenge with a Setup of Speech Enhancement Baseline, *in* 'Procedings of Interspeech 2018', pp. 1571–1575.
URL: *http://www.isca-speech.org/archive/interspeech_2018/chen18d_interspeech.html*

De Wet, F., Badenhorst, J. & Modipa, T. (2016), 'Developing speech resources from parliamentary data for south african english', *Procedia Computer Science* **81**, 45–52.
URL: *https://www.sciencedirect.com/science/article/pii/S1877050916300424*

De Wet, F., De Waal, A. & Van Huyssteen, G. B. (2011), 'Developing a broadband automatic speech recognition system for afrikaans'.

Eiselen, R. & Puttkammer, M. (2014), Developing text resources for ten South African languages, *in* 'Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)', European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 3698–3703.
URL: *http://www.lrec-conf.org/proceedings/lrec2014/pdf/1151_Paper.pdf*

Gillam, R. (1998), 'A rule-based approach to number spellout', *Unicode Consortium [Uni98a], page* .
URL: *https://www.xencraft.com/resources/NumberGeneration.pdf*

Goodman, J. T. (2001), 'A bit of progress in language modeling', *Computer Speech & Language* **15**(4), 403–434.
URL: *https://www.sciencedirect.com/science/article/abs/pii/S0885230801901743*

Groenewald, H. J. & du Plooy, L. (2010), 'Processing parallel text corpora for three South African language pairs in the Autshumato project', *AfLaT 2010* p. 27.
URL: *https://biblio.ugent.be/publication/1851705/file/6736544page=39*

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N. & Wu, Y. (2016), 'Exploring the limits of language modeling', *arXiv preprint arXiv:1602.02410* .
URL: *https://arxiv.org/pdf/1602.02410.pdf*

Louw, J. A., Moodley, A. & Govender, A. (2016), The Speect text-to-speech entry for the Blizzard Challenge 2016, *in* 'Blizzard Challenge Workshop 2016', Cupertino, United States of America.
URL: *http://festvox.org/blizzard/bc2016/MERAKA_Blizzard2016.pdf*

Louw, J. A., Van Niekerk, D. R. & Schlünz, G. I. (2010), Introducing the speect speech synthesis platform, *in* 'Blizzard Challenge Workshop'.
URL: *http://www.festvox.org/blizzard/bc2010/MERAKA$_B$lizzard*2010*.pdf*

McKellar, C. A. & Puttkammer, M. J. (2020), 'Dataset for comparable evaluation of machine translation between 11 south african languages', *Data in brief* **29**, 105146.
URL: *https://www.sciencedirect.com/science/article/pii/S2352340920300408*

Povey, D. (2018*a*), 'Kaldi switchboard rnnlm tdnn-lstm backward example recipe'. Availible online: `https://github.com/kaldi-asr/kaldi/blob/master/egs/swbd/s5c/local/rnnlm/tuning/run_tdnn_lstm_back_1e.sh`, (accessed on 14 May 2022).

Povey, D. (2018*b*), 'Kaldi switchboard rnnlm tdnn-lstm example recipe'. Availible online: `https://github.com/kaldi-asr/kaldi/blob/master/egs/swbd/s5c/local/rnnlm/tuning/run_tdnn_lstm_1e.sh`, (accessed on 14 May 2022).

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P. et al. (2011), The Kaldi speech recognition toolkit, *in* 'IEEE 2011 workshop on Automatic Speech Recognition and Understanding (ASRU)', number EPFL-CONF-192584, Hilton Waikoloa Village, Big Island, Hawaii.
URL: *http://ieeexplore.ieee.org/document/8683713*

Szaszáak, G. & Pierucci, P. (2019), A comparative analysis of domain adaptation techniques for recognition of accented speech, *in* '2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)', IEEE, pp. 259–264.
URL: *http://ieeexplore.ieee.org/document/9089928*

Wilkinson, N. & Niesler, T. (2021), A hybrid cnn-bilstm voice activity detector, *in* 'ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 6803–6807.
URL: *http://ieeexplore.ieee.org/document/9415081*

Xu, H., Chen, T., Gao, D., Wang, Y., Li, K., Goel, N., Carmiel, Y., Povey, D. & Khudanpur, S. (2018), A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition, *in* '2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)', IEEE, pp. 5929–5933.
URL: *https://ieeexplore.ieee.org/document/8461974*

Xu, H., Li, K., Wang, Y., Wang, J., Kang, S., Chen, X., Povey, D. & Khudanpur, S. (2018), Neural network language modeling with letter-based features and importance sampling, *in* '2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)', IEEE, pp. 6109–6113.
URL: *https://ieeexplore.ieee.org/document/8461704*

Yang, X., Li, J. & Zhou, X. (2018), 'A novel pyramidal-fsmn architecture with lattice-free mmi for speech recognition', *arXiv preprint arXiv:1810.11352* .
URL: *http://arxiv.org/pdf/1810.11352.pdf*