Investigating the feasibility of harvesting broadcast speech data to develop resources for South African 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, South Africa
fdw@sun.ac.za

Abstract

Sufficient target language data remains an important factor in the development of automatic speech recognition (ASR) systems. For instance, the substantial improvement in acoustic modelling that deep architectures have recently achieved for well-resourced languages requires vast amounts of speech data. Moreover, the acoustic models in state-of-the-art ASR systems that generalise well across different domains are usually trained on various corpora, not just one or two. Diverse corpora containing hundreds of hours of speech data are not available for resource limited languages. In this paper, we investigate the feasibility of creating additional speech resources for the official languages of South Africa by employing a semi-automatic data harvesting procedure. Factorised time-delay neural network models were used to generate phone-level transcriptions of speech data harvested from different domains.

Keywords: low-resource languages, data harvesting, automatic speech recognition, TDNN-F, domain adaptation

1 Introduction

At least 10 of South Africa's official languages are currently regarded as resource-constrained. New ways to rapidly expand the resources required for ASR development are needed to unlock the full potential of applying state-of-the-art modelling techniques to these languages. Radio broadcasts represent one such potentially unlimited speech data source, which in South Africa is still largely untapped. Broadcasts typically contain various data types, including speech produced by different speakers in various speaking styles but also many non-speech events. For this reason, a data harvesting project was recently initiated in collaboration with the South African Centre for Digital Language Resources[1] to collect, curate and develop new corpora of speech data from broadcast data.

Speech data harvesting could be automated if automatic transcription systems existed. However, the level of ASR technology development in South Africa still has a long way to go before that would be a feasible option for data harvesting in the country. As a first step toward automatic annotation, initial acoustic models to bootstrap the harvesting work were derived from the existing, though limited, NCHLT Speech corpus (Barnard et al. 2014). The viability of using this data set to train deep neural network-based acoustic models was investigated in a previous study in which baseline systems for all 11 official languages were implemented using factorised time-delay neural networks (TDNN-F) (Badenhorst & de Wet 2019). This paper subsequently reports on the adaptation of these models using harvested speech data to improve the quality of the transcription system. The aim is to enhance the system's ability to automatically transcribe speech from various domains.

In a recent study Szaszáak & Pierucci (2019) investigated several approaches to adapt English TDNN-based acoustic models to Indian accented English. They tested different strategies including attempts to employ transfer learning, simple re-training of the entire network for a few epochs and applying i-vector adaptation in Kaldi (Povey et al. 2011). Apart from the speaking style, other differences between existing training data and newly obtained speech examples are important to consider. Regarding i-vector adaptation, for example, Vaněk et al. (2019)

found that a new Czech telephone model trained on short split utterances resulted in poor recognition of much longer real call centre utterances. Further analysis revealed the estimation of i-vector statistics with the TDNN models to be incorrect, resulting in failed adaptation. Conversely, i-vector adaptation in conjunction with data perturbation techniques have been shown to be an effective approach to improving TDNN-based acoustic models for low-resourced languages (Kumar & Aggarwal 2020).

Section 2 of this paper describes the data collection process and its associated challenges. Section 2.1 explains how new data were selected followed by a description of the test data sets that were defined for system evaluation in Section 3. Two of the 11 official languages were included in our investigation: Afrikaans (Afr) and Tshivenda (Ven). Section 4 introduces the important configurations of refined transcription systems, while Section 5 describes procedures to extract shorter segments of speech from raw data. Section 6 provides results for different system configurations. A discussion on the results (Section 7) and subsequent conclusions (Section 8), concludes the paper.

2 Data collection

Initially, it was anticipated that broadcast data from various radio stations would be streamed to a server. 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. A short description of these three data capturing scenarios are as follows:

1. Streaming data from the cloud to local storage
2. Transferring data from individual contractors
3. Transferring data from a hosting service.

In addition, various broadcast data sets could be obtained from each data capturing scenario. For example, some broadcasts contain accurately scripted news bulletins of high quality audio, but only for a limited number of news readers. Scripted news broadcasts were only available under data capturing scenario 2. There are also much larger quantities of unscripted audio data available. South African hosting services provide numerous radio show podcasts in all languages. Collection of broadcasts from individual contractors were severely impacted by the Covid-19 pandemic. Operations at smaller radio stations were diminished as the stations faced many new challenges. One university-based station even had to withdraw from the project. Agreeing on data sharing agreements was also an extremely slow and time-consuming process.

Fortunately, the collaboration with Iono[2] was successful. Agreements could be reached with a few radio stations producing and distributing podcasts on the Iono online audio platform. As a result, the resources collected for the two languages considered in this study could only be obtained under scenarios 1 and 3.

2.1 Content selection

In terms of the speech resources that are available for South Africa's languages, Afr is in a slightly better position than some of the other languages. It was included in the study to develop the harvesting strategy and to provide a broader perspective on different possibilities, given the large variety of podcasts. Ven was chosen as an example of a severely resource-constrained language. Podcasts were much more limited and no previously compiled broadcast news data were available as in the case of Afr. The material from which data were selected for harvesting included three speaking styles: read speech (e.g. news bulletins), conversational speech (e.g. radio dramas), and studio speeches (single speaker messages delivered in a studio).

Factors that were considered to select broadcasts for harvesting included availability, speaking style and the presence or absence of non-speech events like music, advertisements, etc. We also tried to limit the acoustic mismatch between the data to be harvested and the NCHLT data from which the initial

acoustic models were derived. The NCHLT data consisted of clear, read speech prompts with minimal background noise. Therefore, read speech data such as the news or studio speeches were considered first.

For Afr, ample data were available for studio quality speeches. One radio show had more than 85 hours of clear recordings, a single speaker per episode and a sufficiently large number of speakers across all episodes. Speaker meta-data were also available for the episodes. Importantly, the podcasts of this show did not include any start or end jingles, and the duration of an episode was under five minutes, which considerably reduces the need for pre-processing and segmentation. The data represented a new speaking style of clear speech, which the presenter presented in a conversational manner. Data from this radio show was therefore considered suitable for testing NCHLT bootstrapping.

Unfortunately, no show of matching audio quality could be identified for Ven. A set of 1 118 Ven news bulletins that were recorded from streamed radio was therefore chosen as the best available match for the NCHLT data. Each recording included a news bulletin of approximately five minutes, which translates to almost 90 hours of data. The audio also included clips by reporters and news chimes.

3 Test data

Representative test data were required to evaluate the performance of baseline transcription systems as well as the impact of acoustic model adaptation to improve data harvesting for both Afr and Ven. Test sets were therefore selected from the available data, ensuring that the acoustic conditions and speaking styles that occurred in the data selected for harvesting were also represented in the test data. In addition to the existing Afr News test set, four test sets were transcribed manually: Ven News, Afr Messages as well as Ven and Afr Drama. The drama episodes were included because they contain speech produced by multiple speakers, mostly in a conversational style and in various acoustic conditions. The duration in hours and speaker distribution for each test data set is provided in Table 1.

Table 1: Duration and speaker information of test data

Test set	Dur (h)	Spk info
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

4 System refinement

Refined transcription systems were created to enable automatic transcription of the selected content introduced in Section 2.1. In this paper, we report results for the following three types of systems: 1) Baseline, 2) Text-based refined and 3) Acoustically adapted systems. Baseline systems utilised the initial NCHLT TDNN-F models (Badenhorst & de Wet 2019) in combination with a flat ARPA language model consisting of equiprobable phone uni-grams. Text-based refined systems employed language modelling given the limited text corpora that are available in the languages and acoustically adapted systems were built by updating the acoustic models within the above configurations using automatically transcribed data. The Kaldi toolkit was used to create all the acoustic models that were used in this study. It was also used to perform segmentation and acoustic model adaptation.

4.1 Text-based refinement

To configure the text-based refined transcription systems, language models were built from a number of text corpora produced during various previous projects for all 11 languages. Applying these texts to system development ensured that the systems developed for different languages would be comparable, since the same types of texts are available in each language. TTS prompts originating from the CSIR Lwazi projects proved most useful for configuring the baseline transcription systems. Subsequent

combinations of the TTS prompts with text from the CSIR NCHLT Speech[3], CTexT NCHLT Text (Puttkammer et al. 2014*a*,*b*), and CTexT Autshumato (McKellar & Puttkammer 2020, Groenewald & du Plooy 2010) projects were also evaluated. Table 2 presents the unique word token counts (N=1) for each text in Afr and Ven. In addition, the counts of word 2-grams (N=2) provides some perspective on word sequences and the total number of tokens (T) in each text report on the size of the corpora.

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 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 contained only 3 488 unique words, fewer words than the NCHLT speech data annotations. While the NCHLT Text corpus contained more than seven times the vocabulary of the Lwazi TTS text, compared to the Afr component, the vocabulary size was about half the number of unique words in Afr. The total number of Ven NCHLT Text corpus words was approximately 1 million. However, the larger texts may include more out-of-language words than the small, curated TTS text resources. Proper names and English terminology that occur in the NCHLT Ven text could contribute to the larger vocabulary sizes.

ASR systems with vocabulary derived from limited text cannot predict out-of-vocabulary (OOV) words in the test data. To better understand the impact of OOV words on text-based refined system results, Table 3 summarises the OOV rates for the TTS, NCHLT Speech, NCHLT Text as well as a combination of the NCHLT Text and TTS texts (Txt Corp + TTS). It is clear that predicting news data using the limited TTS text (especially in the case of Ven) would include more errors due to the larger OOV rate of these data sets. The values indicate far fewer OOV words for the larger NCHLT Text corpus. Moreover, the vocabulary of these texts produces similarly low OOV rates for both news and drama test data sets.

4.2 Acoustic refinement

The different speaking styles in broadcast data necessitates refinement of the acoustic models. In this work, the adaptation approach was based on re-taining the initial NCHLT model, but keeping NCHLT i-vectors intact. This meant that i-vectors for the adaptation data were also generated using the NCHLT i-vector extractor.

The approach featured a two-stage TDNN-F model adaptation recipe. For adaptation data, we used the automatic transcriptions of 6-gram text-based refined systems generated for the new data. The first stage of the adaptation recipe reproduced the training setup of the initial model, but instead of finalising the process left the training setup in such a condition that a second stage of training could be applied. The second stage is therefore an adaptation stage, re-training the standard (four-epoch) model for an additional epoch, now using the adaptation data. To enable model training and standard feature extraction, triphone alignments and data perturbation was applied to both the NCHLT and adaptation data. The initial NCHLT models were used to perform triphone alignment.

With re-training the intensity of the adaptive training was controlled, adjusting the learning rate for the training iterations of the last epoch and not the number of iterations (or training for more epochs). In essence, the standard start and end learning rate thresholds of the TDNN-F recipe across the initial four epochs of training were left to the same setting, restricting the training algorithm to these limits. For training of the last epoch, similar thresholds to the lower threshold of training (0.000020 and 0.000015 respectively) were applied. This restricted the algorithm to apply a learning rate close to the lower setting for the remaining iterations of training.

Table 2: Comparison of vocabulary sizes for text data sources

		Lwazi	NCHLT		Autshumato			
Language	**N**	**TTS**	**Speech**	**Text corpus**	**Words**	**Phrases**	**Eval**	**Parallel text**
Afr	1	12 447	8 565	56 192	11 785	1 226	3 015	30 440
	2	66 652	18 205	424 438	61	1 061	11 126	167 830
	T	**143 958**	**173 128**	**2 357 560**	**11 892**	**2 508**	**38 125**	**2 341 627**
Ven	1	3 488	7 578	24 314	4 435	994	2 877	-
	2	12 134	26 135	198 136	62	1 636	15 132	-
	T	**30 835**	**217 526**	**996 393**	**4 899**	**3 894**	**45 513**	-

Table 3: Out-of-vocabulary word rates (as a percentage) of different texts given the test data

	Lwazi	NCHLT		
Test set	**TTS**	**Speech**	**Txt Corp**	**Txt Corp + TTS**
Afr News	18.46	20.69	6.96	6.38
Afr Messages	6.28	17.48	3.83	2.86
Afr Drama	11.23	19.02	6.69	6.11
Ven News	21.66	14.50	6.44	6.44
Ven Drama	17.57	13.62	8.22	8.09

5 Data harvesting

As said in Section 2.1, data harvesting usually requires audio segmentation. ASR systems are trained on relatively short segments of speech, excluding non-speech events such as jingles and music. In this study, two different segmentation techniques were applied. The first, a speaker diarisation-based approach, enabled automatic detection of the start of the news inside the 10 minute Ven recordings. A second Kaldi alignment-based silence detection method was then applied to both the Afr and Ven harvest data. Sufficient segmentation was possible for detected silence labels of 0.1 seconds or longer in duration. The produced segments had durations between 5 and 15 seconds.

5.1 Segmentation

To detect the start of news (after the news chime), we applied a heuristic algorithm. The heuristic utilised an unsupervised implementation of speaker diarisation, as implemented in the Open-Source Python library for audio signal analysis (Giannakopoulos 2015). Applying the speaker diarisation to each recording and setting the predefined number of clusters to three and four speaker labels, respectively, resulted in separable classes. The other chosen parameter values of the heuristic were informed by a short analysis previously conducted on the manually segmented Ven news test set segments. As reported in Table 1, the Ven News reader segments had a total duration of almost 33 minutes and the mean and median duration of the segments was about 38 seconds. This meant that the typical news reader segment may be about 40 seconds or longer. Subsequently, the speaker cluster labels of the new recordings were validated to check if such long segments were detected within the first 120 seconds of each recording. For 44% of the recordings, no detections were made, so the minimum segment duration was lowered to 20 seconds instead. This lowered the non-detection rate further to 12% of the recordings. Given these findings, the following heuristic steps were implemented to select the first news segment at the beginning of the recording:

1. Select only those segments starting at least two seconds after the start of the recording, since inspection revealed that longer segments starting at time zero seconds usually contained music.
2. Select only segments with a minimum duration of 25 seconds.

3. Compute the overlap in time between segments created by the three- and four-speaker label diarisation for the first 180 seconds of audio.
4. For any overlapping segments detected during Step 3, set the start time to that of the first segment with an overlap of 50% or more.
5. Cut the audio files so that each file starts at the updated start time.

Applying the steps above ensured that for 90% of the recordings news start times could be detected. Spot checks confirmed valid news starts.

5.2 Segment selection

The complete Ven bulletin recordings included news clips, some of which contained English or another language. Furthermore, some or part of the news chimes might be included in a segment since the news not only started with, but also ended with the chime and music. Apart from these factors, the first automatic segment transcriptions would include labelling errors. In fact, the previous study that led to the development of the initial Afr acoustic model highlighted the importance of acoustic selection for imperfect speech training data, applying phone-based dynamic programming (PDP) scoring. Therefore, the adaptation experiments performed in this study were conducted by first including those audio segments that showed the best PDP scores given the new broadcast domains. To accomplish this, subset selections of the adaptation data were made. The PDP scoring technique requires two transcriptions of the same segment of audio. One transcription is usually produced by free phone recognition (employing a language model based purely on a string of phone labels). The second transcription comes from a better, more refined recognition system.

To create new speech corpora from the selected data (Section 2.1), the best automatic transcription systems were applied to transcribe each of the audio segments to be harvested. For each segment, a final PDP score was also estimated. In this study, the final refined systems were sometimes based on word recognition (see Figure 2) whenever lower error rates could be obtained this way.

Table 4: Baseline and better text refined systems PERs

Test set	Flat ARPA	6-gram ARPA
Afr News	20.98	16.34
Afr Messages	25.82	22.42
Afr Drama	42.69	39.52
Ven News	27.12	20.84
Ven Drama	45.80	40.35

6 Results

To test transcription performance, each of the test data sets were transcribed by the transcription systems described in Section 4.1. Subsequent phone error rates (PERs) were calculated including only speech phone labels during the estimation. Table 4 shows the performance achieved for two types of systems: 1) flat phone-based ARPA language model transcription and 2) systems applying 6-gram phone-based ARPA language models given the text data. In an iterative strategy, different ARPA-based systems were built for combinations of the text data. The PER values in Table 4 reflect the best 6-gram systems in combination with NCHLT acoustic models. Only the best 6-gram text refined systems in each language was chosen. This meant that the 6-gram language model was derived from the TTS text for the Afr systems and from a combination of the TTS and NCHLT Text corpus in the case of Ven.

Utilising the 6-gram language models, systems produced substantially lower PERs for all Afr and Ven test data sets. Similarly, results for a combination of the TTS and NCHLT Text corpus are presented for Ven. The results show higher PERs (above 40%) for Drama data compared to News data. In each case, the 6-gram ARPA systems deliver an improvement over just using a flat ARPA transcription system.

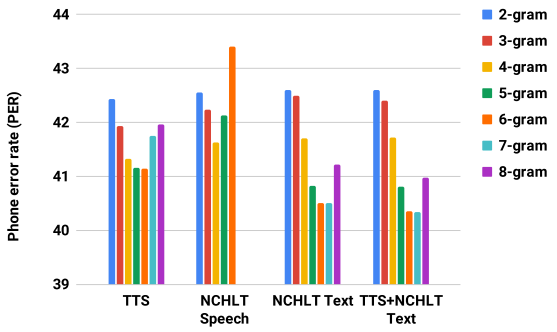

Figure 1: Ven PERs on drama data for various texts and context sizes

The 6-gram context size of the language models were chosen as a point where lower PER rates could be achieved in both languages for TTS and larger sets of text data. Figure 1 presents an example of such an analysis. Lower error rates were measured as the context size of the n-gram language models increased up to a point of about 6 or 7-gram phone contexts. Except for the NCHLT Speech text, this finding generalised well for the various combinations of text.

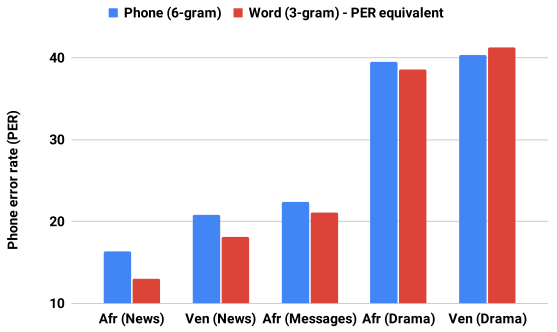

Figure 2: Comparing PER equivalents for phone and word recognition systems

Interestingly, even lower PERs could be obtained for most test sets when applying word recognition first. Using 3-gram language models, the automatic word-level transcriptions were converted to phone labels using a pronunciation dictionary. From the arrangement in Figure 2 it seems that the phone transcriptions for test sets with better transcription rates, such as those for the News and Messages data, benefited most from recognition at the word level. Only for the Ven Drama test data with a PER of over 40% word recognition did not provide a benefit.

Table 5: Comparing the PERs between NCHLT baseline and adaptively trained models on Afr Messages episodes.

	Flat ARPA		6-gram ARPA	
Test set	**NCHLT**	**Adapt**	**NCHLT**	**Adapt**
Afr News	20.98	21.25	16.34	16.56
Afr Messages	25.28	21.92	22.42	17.71
Afr Drama	42.69	43.70	39.52	39.87

Subsequently, a first iteration of acoustic refinement was applied to evaluate the potential benefit of acoustic model adaptation. Table 5 shows the effect that using the adapted acoustic models in conjunction with the phone language models had on recognition. In this experiment, almost the entire Afr Messages data set (71 hours, excluding the test data) was applied as adaptation data, treating each 3 minute episode as a single audio segment. Clearly, PERs were reduced more for the Afr Messages test data from the same speaking style. These results compared well with what was also produced using about 18 hours of adaptation data, if PDP scoring was applied to select shorter segments of well-transcribed data from the larger set for adaptive training first.

7 Discussion

Language models built from the vocabulary of the large corpora and combinations of texts produced fairly low OOV rates. The values in Table 3 showed that OOV rates of less than 10% across the test data sets representing different speaking styles from different sources (news, studio messages and drama episodes) could be achieved. While this fact may seem promising in terms of vocabulary coverage, PERs applying these language models did not reflect such significant reductions when employing the larger texts. Instead, fairly low PERs could be

obtained using phone language models with adequate phone coverage, such as the models based on phonetically balanced TTS prompts alone. Only small improvements were seen for Ven where the size of the TTS prompts text was much smaller than that of the Afr. The analysis in Figure 1, where Ven drama transcriptions were analysed provides an example. It records less than 1% PER difference for employing additional text beyond the TTS text. Therefore, including more vocabulary from additional text corpora (other than those already included) next, might not benefit the construction of phone level systems built with baseline acoustic models as much as first thought. Transcribing the conversational speaking style test sets generated significantly more error than news reading test data. For continued data harvesting, the fairly high error suggests that more acoustic development is required to lower PERs, before building larger vocabulary language models. Another perspective on transcription error comes from analysis of the context size of the phone language models. Larger phone contexts played a significant role to lower the PER.

It was also shown that with larger context phone language models, larger reductions in the PER were achieved for the test sets where lower PERs had been obtained in the first place. This finding supports the idea that larger vocabulary systems require better (well-matched) acoustic models. In essence, the usefulness of including the information of text resources through language modelling in transcription systems increases as the acoustic modelling improves. Initial word recognition tests produced a relatively high transcription error for the test data. As with well-resourced languages, where acoustic models can be developed with sufficient audio data, incorporating larger texts should, in future, make a significant contribution to achieve more accurate recognition.

The adaptation strategy that was employed successfully produced significant PER reductions, adapting to the domain of the new Afr studio messages target data set. The speakers in the selected radio show spoke in a more conversational manner. A similar observation was made for the Ven language, where adapting to the news domain also produced some improvement but, as expected, this adaptation did not really improve Ven drama recognition. These findings proved the ability of the adaptation technique to adapt and develop TDNN-F acoustic models for new acoustic domains.

More surprisingly, these adaptations could also be achieved utilising less than 20 hours of data, choosing the best ranking PDP scored audio segments. While it is expected that the technique will produce a similar outcome when applied to data from other languages, future work should still focus on iterative training using the updated transcriptions that are produced. Given the unbalanced ratio of accurately-transcribed NCHLT training data to the automatically transcribed adaptation data samples, it was also necessary to keep using NCHLT i-vectors created by a speaker-specific transformation that is required for TDNN-F training. Additionally, this strategy has the advantage of being speaker independent and does not require speaker labels for the adaptation data. As more accurately transcribed adaptation data samples become available, it should be adjusted to include appropriate i-vector adaptation.

8 Conclusion

The results of this study indicate that the effectiveness of including text data as a language resource during automatic transcription system development depends on the state of the acoustic models employed for harvesting. While including a well-balanced within-language text (such as a set of TTS prompts) is important, acoustic models that are well-matched to the acoustic domain of the speech data to be harvested, are also required. Including larger quantities of text for language modelling then becomes more effective. While the above approach would create more data, reducing the transcription error would increase the yield of correctly-annotated speech data. It is recommended that new ways to improve the model development should be sought. Better acoustic adaptation should eventually be possible for sets of

adaptation data with sufficiently accurate transcriptions.

The outcome of this study proved that the current data harvesting technique would be applicable to all official languages of South Africa. Good generalisation to speech data from a new domain, incorporating a new speaking style, was achieved for Afr. Secondly, the same approach could be duplicated successfully in the much more resource-constrained Ven language, producing PERs on broadcast news data that were similar to the PER obtained for the Afr data from the new speaking style.

Notes

[1] `https://www.sadilar.org/`, SADiLaR is funded by the South African government's Department of Science and Innovation (DSI).

[2] `https://https://iono.fm/` Iono maintains an online audio platform providing podcasts and audio live streaming services for a variety of sources.

[3] NCHLT Speech text comprised 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).

References

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*

Barnard, E., Davel, M. H., van Heerden, C., de Wet, F. & Badenhorst, J. (2014), The NCHLT speech corpus of the South African languages, *in* 'in Proc. SLTU', St Petersburg, Russia.

Giannakopoulos, T. (2015), 'pyaudioanalysis: An open-source python library for audio signal analysis', *PloS one* **10**(12).

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.

Kumar, A. & Aggarwal, R. K. (2020), 'Hindi speech recognition using time delay neural network acoustic modeling with i-vector adaptation', *International Journal of Speech Technology* pp. 1–12.

McKellar, C. A. & Puttkammer, M. J. (2020), 'Dataset for comparable evaluation of machine translation between 11 south african languages', *Data in brief* **29**, 105146.

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.

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

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

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.

Vaněk, J., Michálek, J. & Psutka, J. (2019), Tuning of acoustic modeling and adaptation technique for a real speech recognition task, *in* 'International Conference on Statistical Language and Speech Processing', Springer, pp. 235–245.