Unsupervised acoustic model training: comparing South African English and isiZulu

Neil Kleynhans[1], *Febe de Wet*[2] *and Etienne Barnard*[3]

[1,3] Multilingual Speech Technologies, North-West University, Vanderbijlpark, South Africa
[2] Human Language Technologies Research Group, Meraka Institute, CSIR, South Africa
[2]Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
{ntkleynhans,fdwet,etienne.barnard}@gmail.com

Abstract

Large amounts of untranscribed audio data are generated every day. These audio resources can be used to develop robust acoustic models that can be used in a variety of speech-based systems. Manually transcribing this data is resource intensive and requires funding, time and expertise. Lightly-supervised training techniques, however, provide a means to rapidly transcribe audio, thus reducing the initial resource investment to begin the modelling process.

Our findings suggest that the lightly-supervised training technique works well for English but when moving to an agglutinative language, such as isiZulu, the process fails to achieve the performance seen for English. Additionally, phone-based performances are significantly worse when compared to an approach using word-based language models. These results indicate a strong dependence on large or well-matched text resources for lightly-supervised training techniques.

Index Terms: lightly-supervised training, unsupervised training, automatic transcription generation, audio harvesting, English, isiZulu

1. Introduction

Vast amounts of audio data are created on a daily basis. Typical sources are radio / television broadcasts, podcasts and lectures. Very few of these audio corpora have corresponding orthographic or other transcriptions. A particularly interesting scenario where large amounts of audio data are created and where text transcriptions would be of great benefit are call-centre environments. Access to text representations of the audio would aid in swiftly analysing the data and making necessary adjustments where appropriate.

Manually transcribing audio data is a resource intensive process requiring disproportionate amounts of money and time. The time to produce a transcription depends on transcriber expertise and required accuracy of transcription: Approximate transcriptions can be generated at 3 to 5 times real time while for highly accurate transcriptions the time considerably increases to 50 times real time [1]. This turnaround time is often too long to make business sense, considering the amount of audio data collected in a day, and so automatic means become the only feasible route.

An automatic solution would require an automatic speech recognition (ASR) system at its core, but in general one would not have access to matching in-domain audio and text data. To sidestep this dilemma, the "lightly supervised" acoustic model (AM) training approach [2] provides a mechanism to develop and constantly refine AMs needed by an ASR system. This approach, however, requires approximate transcriptions (for instance, closed-caption transcriptions are frequently employed). Another approach is "unsupervised" AM training [2, 3] which follows the same broad steps as the lightly-supervised approach but does not make use of approximate transcriptions. What makes these approaches attractive is the minimal initial resource investment. An added incentive is that the AMs are trained on in-domain data, which removes the mismatch between the audio data and the AMs.

In this paper we investigate factors related to the resources required to start and maintain the automatic harvesting of untranscribed audio data using an unsupervised AM training approach. Specifically, we investigate the scenario for the resources-constrained South African English (SAE) and isiZulu languages.

2. Background

Lightly-supervised and unsupervised acoustic model training have been applied in many different scenarios [2, 3, 4, 5, 6, 7, 8, 9]. The basic steps of the iterative algorithm are [2] the following:

- Partition the audio data into homogeneous portions based on characteristics such as channel, environment or speaker.
- Normalise all approximate transcriptions (if available) and produce appropriate phonetic pronunciations.
- Develop seed acoustic models by manually transcribing a small portion of the in-domain audio data or using existing AMs.
- Automatically transcribe the raw audio data using well-trained or biased language models (LM) or word-graphs trained on the approximate transcriptions (e.g. closed captions).
- As an optional step, use acoustic models to align the audio and the available approximate transcriptions. Remove audio data where the alignment and automatic transcription disagree excessively.
- Use the newly transcribed data to re-train AMs and repeat the process.

If the penultimate step is not implemented then the lightly-supervised approach collapses to an unsupervised training approach.

Lamel *et al.* [2] investigated the minimal requirements needed to bootstrap the lightly-supervised and unsupervised

processes. AMs trained on 10 minutes of audio data and a LM developed on 1.8 M word corpus (40k lexicon) were sufficient to initiate the process. On 200 hours of raw audio data, using the Topic Detection and Tracking (TDT-2) English corpus, the minimal system obtained a final word-error-rate (WER) of 28.8%, which was significantly worse compared to 18% achieved by a system utilising a LM trained on text corpora (Hub4 and TDT corpora) – but these corpora contain orders of magnitude more text. In a completely unsupervised approach, the minimal system achieved a WER of 37.4% on 140 hours of raw audio. Similarly, the same unsupervised approach was applied to train AMs on Portuguese broadcast news audio [4]. The 3.5 hour trained AM produced a WER of 42.6% while AMs trained on automatically transcribed 30 hours delivered a WER of 39.1%. The LMs were trained on news-related text corpora containing 72.6M words.

Novotney *et. al.* [5] experimented with limited amounts of labelled audio and text data in their unsupervised AM training investigations. They focused on the Fisher corpus data, containing telephone-quality conversational speech and had access to a text corpus containing 1.1 billion words. Interestingly, their findings suggest weaker LMs do not severely impact the unsupervised training of AMs and have a greater impact when decoding the actual evaluation set.

To improve their transcribing system Nguyen and Xiang [6] added 702 hours of audio selected from a 1 400 hour audio data set using the lightly-supervised procedure. The audio data was selected from the TDT English corpora – TDT2, TDT3 and TDT4. The initial acoustic models were trained on 141 hours of audio data and the subset-specific (depending on data set) LMs were trained by interpolating from a 360M word common LM. The baseline system WER of 12.1% was reduced to 10.1% using the 702 hours of added training data.

Gales *et. al.* [8] made use of lightly-supervised training to automatically transcribe audio data from the TDT and 2003 BN collection corpora. The biased LMs were trained on broadcast news text as well as closed captions. The LM weighting was similar to that of Hazen [1] – 90% closed-caption text and 10% general broadcast news.

Chan and Woodland [7] applied lightly-supervised training to 500 hours TDT2 and 300 hours TDT4 corpora. The LMs were trained on text sourced from the closed captions as well as closely related text. The entire text corpus consisted of approximately a billion words. The out-of-vocabulary rates were 0.68% and 0.47% for the different corpora. Again, a biased LM was used with interpolation weights of 0.92 and 0.90, respectively.

Gollan *et. al.* [3] utilised unsupervised training to improve upon baseline AMs trained on 100 hours of manually transcribed audio data selected from English-only European Parliament Plenary Session speeches. Adding 180 hours of automatically transcribed audio improved the WERs from 10.4% to 9.6%.

Davel *et. al.* [9] showed that lightly-supervised automatic harvesting for ASR resource creation in a resource-scarce environment does not require well-trained LMs. In their approach, a phone-based ASR system was used to automatically generate transcriptions, for roughly 100 hours of SAE radio broadcast audio data, using a flat-phone task grammar. The seed models were initially developed on US English data and gradually replaced by the in-domain SAE dialect. Data filtering was achieved by using a garbage model that absorbed badly aligned audio portions.

Gelas *et. al.* [10] made use of a Swahili ASR system to aid in speeding up the task of manually transcribing a 12 hour audio portion of a 200 hour Swahili web broadcast news speech corpus. The initial ASR system was trained on 3.5 hour read speech Swahili corpus. The procedure used an ASR to automatically transcribe a 2 hour portion of audio. These transcriptions were manually corrected. After the correction process, the newly transcribed audio data was used to increase the amount of training data used to train the new AMs. The process was repeated until 12 hours were transcribed. Utilising the ASR system to automatically transcribe the data reduced the manual correction time from an initial 40 hours to a final 15 hours. The LM was trained on a text corpus which contained 28 M words and had a 65k lexicon.

Previous investigations suggest that the unsupervised AM training approach does not require vast resources to begin the harvesting process. As few as 10 minutes of labelled audio data to train AMs and 100k – 1M words to train language models. Some approaches do not require LMs – Davel *et. al.* [9] – but approximate transcriptions were available for data filtering. If LMs are used, however, the text corpora are quite well matched to the domain which in a resource-constrained environment will not be easy to access or develop.

In this study we therefore investigate:

- unsupervised AM training without the aid of a language model – phone decodes only – ,
- the usefulness of language models trained on unrelated text corpora, and,
- the effect of text corpus size used to train N-gram LMs.

3. Method

3.1. Corpora

The NCHLT corpus is a read-speech corpus containing high-bandwidth audio data and transcriptions thereof for all eleven South African languages [11]. Mobile devices were used to collect the audio data. The transcriptions contain short sentences and were derived from large text corpora in order to attain coverage of the most common triphones of the target language. For our unsupervised AM training investigations we limited ourselves to using the English and isiZulu sub-corpora.

3.1.1. NCHLT English

The English NCHLT sub-corpus contains audio data collected from 210 different speakers. There are a total of 77 412 utterances with each speaker contributing roughly 500 utterances. Table 1 shows the duration in hours, the number of speakers and utterance amount for the training and evaluation data sets for the NCHLT English sub-corpus

Table 1: *The duration, amount of speakers and number of utterances for the NCHLT English sub-corpus.*

Data Set	Duration (Hours)	# speakers	# utterances
Training	54.19	202	74180
Evaluation	2.42	8	3232

There is a total of 223 561 tokens and a lexicon of 8 350 words for the entire corpus. The training set contains 214 192 tokens in total and a lexicon of 8 328 words, while the evaluation set contains a total of 9 369 tokens and a lexicon of 3 627 words. The out-of-vocabulary (OOV) rate between the training and evaluation set is 0.61%.

3.1.2. NCHLT isiZulu

Similar to the English sub-corpus, the isiZulu NCHLT sub-corpus contains audio data collected from 210 different speakers. There are 44 673 utterances in total with each speaker contributing roughly 500 utterances. Table 2 shows the duration in hours, the numer of speakers and utterance amount for the training and evaluation data sets for the NCHLT isiZulu sub-corpus

Table 2: *The duration, amount of speakers and number of utterances for the NCHLT isiZulu sub-corpus.*

Data Set	Duration (Hours)	# speakers	# utterances
Training	52.23	202	41871
Evaluation	4.02	8	2802

There is a total of 133 480 tokens and a lexicon of 25 651 words for the entire corpus. The training set contains 125 028 tokens in total and a lexicon of 25 231 words, while the evaluation set contains a total of 8 452 tokens and a lexicon of 5 189 words. The out-of-vocabulary rate between the training and evaluation set is 8.1%.

3.2. Pronunciation Modelling

The pronunciation dictionaries for the NCHLT sub-corpora were sourced from previous work as outlined in Davel and Martirosian [12].

The English pronunciation dictionary contained 15 000 unique entries and a phone set of 43 phones. Phonetisaurus [13] was used to perform grapheme-to-phoneme (G2P) prediction for words not found in the seed pronunciation dictionary. Phonetisaurus implements a WFST-driven G2P framework that can rapidly develop high quality G2P or P2G systems. The English NCHLT text required 3 966 G2P predictions.

For isiZulu a more elaborate approach was followed. For isiZulu words only G2P prediction was performed using the default&refine algorithm proposed in [14], while for code-switched English words, the above Phonetisaurus G2P prediction was used. To identify English words a simple N-gram text-based language identification was implemented. The MIT language modelling toolkit [15] was used to build 3-gram back-off LMs for both English and isiZulu. The training word sets were extracted from the seed pronunciation dictionaries – the English word set had 15 000 words in total and isiZulu had a total of 15 404 words. A word was classified based on the perplexity score. Once all words with missing pronunciations were predicted the English phone set was mapped to the isiZulu phone set using manual rules. Lastly, the isiZulu phone set was further mapped using the MultiPron rules [16] which resulted in 32 phones in total.

3.3. ASR system

The speech recognition system development follows a similar structure to that described in Kim *et. al.* [17]. The audio data was converted to Perceptual Linear Prediction (PLP) coefficients. The 52 dimensional feature vector was created by appending the first, second and third derivatives to the 13 static coefficients (including the 0'th component). Corpus-wide mean and variance normalisation was applied.

AMs were developed by following an iterative training scheme. Firstly, 32-mixture context-independent (CI) AMs were trained and used to produce state aligns for the CI AMs trained in the initial development of cross-word triphone context-dependent (CD) AMs. Once the CD AMs were trained the process was repeated and the previous AMs were used to produce all state alignments before the model mixture incrementing phase. The process was repeated twice for all experiments.

All Hidden Markov Models (HMM) employed a three state left-to-right structure. Each CD HMM's state contained eight mixture diagonal covariance Gaussian models. A question-based tying scheme was followed to create a tied-state data sharing system [18] - where any context-dependent triphone having the same central context could be tied together.

Once the CD AM development was completed, Heteroscedastic Linear Discriminant Analysis (HLDA) was applied to reduce the 52-dimensional PLP feature vectors to a dimension of 39. A global transform was used for the estimation – a single class for all the triphones. After estimating the HLDA transform, the CD AMs' parameters were updated. It was found that allowing the variances to be updated resulted in a large percentage of floored variances. Therefore, only the weights and mean parameters were updated. Two update iterations were performed.

Lastly, Speaker Adaptive Training (SAT) was applied using Constrained Maximum Likelihood Linear Regression (CMLLR) transformations. The same HLDA global transform was used and the CD AMs were updated twice – only weights and means.

The decoding task was a two-step process. The HLDA CD AMs were used to automatically generate transcriptions and a speaker-based CMLLR transform estimated. Then the CMLLR was applied on the second decoding pass.

3.4. Language Models

To investigate the effect of developing LMs on mismatched text corpora, two alternate sources of text unrelated to the NCHLT corpora were used. Before training the LM, the text had to be normalised. This involved,

- Removing punctuation marks.
- Converting numbers to written form.
- Converting characters to lower-case.

Once normalised, the MIT-LM toolkit was used to develop the LMs. Only back-off bigram LMs were created due to limitations of HVite (the HTK decoder). For probability smoothing, fixed Kneser-Ney smoothing was applied.

3.4.1. English

The English LM was developed on a 1.6M word text corpus. The text forms part of the 109M word South African Broadcast News (SABN) text corpus [19]. This corpus contains text extracted from a number of major South African newspapers. The text contain 1 692 929 tokens and a lexicon of 45 664 types. The OOV rate between the SABN and NCHLT text is 29.17%.

3.4.2. isiZulu

The isiZulu LM was developed on a text corpus provided by the Centre for Text Technology (CText) [20]. The original corpus had 223 709 tokens but after applying text normalisation this number increased slightly to 234 216 (due to number expansion). The lexicon associated with the processed text was 38 869 types. The OOV between the CTEXT and NCHLT corpora was 71.59% – the OOV rate is far from ideal and may negatively influence the isiZulu results. Similar Zulu OOV rates

have been seen in the investigation performed by Gales *et. al.* [21] and is common to morphological rich languages.

In addition to training a word-based LM, a syllable LM was also trained. Only words classified as isiZulu, using the text-based N-gram classification approach, were split into syllables. After splitting the words of the LM development text, there were 677 971 tokens and 7 759 types.

3.5. Unsupervised training

To investigate the effectiveness of unsupervised AM training on SAE and isiZulu, the training sets of the NCHLT corpora were partitioned into a number of non-overlapping portions. The seed AMs were trained on a single hour selected at random from the entire training set and from 50 speakers. The transcriptions were used during the seed AM training which simulates the need for manually transcribing a portion of the audio if no other AMs or labelled data is available.

The remainder of the training data set was partitioned into smaller 3, 6, 12 and 24[1] hour sets of untranscribed audio. At each stage, the previously transcribed data sets, including the seed data, were pooled and used to develop new AMs. The current stage's untranscribed data was transcribed using the new AMs set.

To measure the progress of the unsupervised model training and accuracy of the models, phone-error-rates (PER) are reported on the evaluation set as well as the data set that was transcribed. PER are reported since, in the HTK model training recipes, only phone level representations are needed to train acoustic models.

Lastly, two methods of unsupervised acoustic model training were investigated. These are phone-based (flat phone grammars) and word-level LM-based approaches. Additionally, for isiZulu the syllable LMs are also investigated.

4. Results

The PERs of the AMs developed through unsupervised training are reported. The accuracy is measured in terms of automatically transcribing the successive data portion (if for instance the AMs are trained on the seed plus three hours of data, the next six hour data set is viewed as a "testing" set) and the evaluation data set – the successive data portion set is labelled "Raw" and the evaluation set is labelled "Eval".

4.1. English

Table 3 shows the AM PERs for increasing amounts of automatically transcribed data and using a flat phone-based grammar to harvest more data. Interestingly, when adding three hours of automatically transcribed data the error rates on the evaluation set decreases by more than 2% absolute; a 3% absolute decrease is seen on the raw six hour data set. After the +3 hour mark, however, PERs increase as more automatically transcribed data is used to train AMs.

Table 4 shows the PERs of AMs trained on automatically transcribed data and using a word LM model when decoding the data. The general trend is a decreased in PERs as more data is used to train the AMs, which is consistent with trends seen in literature. (The "Raw" +24 hr experiment was not reported in table 4, as "Raw" and "Eval" results are highly correlated and the same performance can be expected for the +24 hr "Raw" case).

[1] All the remaining data was around 24 hours in duration.

Table 3: *The accuracy of the English acoustic models developed using a flat phone grammar approach.*

Data Set	Raw	Eval
Seed (1 hr)	42.47	40.36
+ 3 hr	39.33	38.32
+ 6 hr	39.9	38.85
+ 12 hr	41.55	40.21

Table 4: *The PERs of different English AMs trained on increasing portions of automatically transcribed data using a LM.*

Data Set	Raw	Eval
Seed (1 hr)	25.26	23.52
+ 3 hr	21.33	20.39
+ 6 hr	19.11	18.74
+ 12 hr	14.66	14.73
+ 24 hr	-	13.98

4.2. Text data dependency for English

Novotney *et. al.* [5] suggested that the size of the LM has a limited effect on the unsupervised training of AMs. To investigate this, we limited the amount of text used to train the English LMs. The text was limited to half and then a quarter of the full text. Table 5 shows the number of tokens, types and OOV rate for various sized text corpora used to train different LMs.

Table 5: *Tokens, types and OOV of text used to develop LMs on full, half and quarter amounts of the English text corpus.*

Percentage of Full Text	Tokens	Types	OOV
100 %	1.69 M	45k	29%
50 %	846k	35k	34%
25 %	423k	26k	40%

Table 6 shows the AMs correctness and accuracies developed on increasing portions of automatically transcribed data using a LM trained of half the text corpus. As with the full text trained LM, all values increase as the amount of automatically transcribed data is used to train the AMs.

Table 7 shows the PERs obtained by using various AMs trained on automatically transcribed data and using a LM trained on a quarter of the full text corpus. Again, decreasing trends can be seen.

Considering the final evaluation results for AMs trained on all the acoustic data and the quarter, half and full sized LMs (17.36%, 16.78% and 13.98%), we can see a slight increase in accuracy as more text data is used to develop the LM. The drop in performance may also be attributed to the increase in OOV rates, observed for the LMs trained on less text data.

4.3. isiZulu

Table 8 shows the performance of AMs trained on increasing amounts of automatically transcribed data using a flat phone decoding grammar. Besides a slight decrease in error rate for the raw set at the added three hour mark, the remaining PERs for both data sets steadily increase as more automatically transcribed data is added to the training pool.

Table 9 shows the PERs for various AMs trained on increas-

Table 6: *The performance of English AMs used to automatically transcribe data using a LM developed on half the available text data.*

Data Set	Raw	Eval
Seed (1 hr)	25.7	23.92
+ 3 hr	21.95	20.81
+ 6 hr	19.51	18.9
+ 12 hr	17.95	17.6
+ 24 hr	-	16.78

Table 7: *The performance of English AMs used to automatically transcribe data using a LM developed on a quarter of the available text data.*

Data Set	Raw	Eval
Seed (1 hr)	25.7	24.99
+ 3 hr	22.4	21.49
+ 6 hr	20.07	19.44
+ 12 hr	18.69	18.31
+ 24 hr	-	17.36

ing amounts of harvested data and using a word LM during the decoding process. As with the flat phone approach, the same overall increasing trends are observed, however, the absolute performance values are somewhat lower.

Table 10 captures the performances of the AMs trained in an unsupervised manner while using a syllable LM during decoding. Again, there is a general increasing trend in the PERs as more automatically transcribed data is used to develop the AMs – except for the evaluation PER which decreases slightly for the added three hour mark. The PER values of the syllable approach are consistently better compared to the flat phone approach, but in general worse compared to the LM-based approach.

To try and rule out the possibility of poorly trained seed models, a different seed model trained on three hours of data was tried. Table 11 shows the PER percentages for AMs trained on increasing amounts of automatically transcribed data using a word LM during the decode cycle. Compared to the single hour seed model, the performance measures are slightly better but again the same increasing trend is observed as more data is added to the training pool.

5. Conclusion

In this study we applied the well-known unsupervised acoustic model training scheme to resource-scarce South African English and isiZulu audio data. We investigated phone-based and word-based language models and, in addition, a syllable language model for isiZulu. The default seed acoustic model was trained on a single hour of manually transcribed data. The text corpora used to develop the language models were selected from unrelated sources which differed significantly in the OOV rates – 29% and 76% for English and isiZulu respectively. For English, we also experimented with the amount of text data used to train the language model.

From our results we may conclude:

- The unsupervised acoustic model training scheme performs well for SAE if a word-based LM is used.
- The phone-based approach, for English and isiZulu, did not achieve increasingly better results as more automatically transcribed data was added to the training pool.

Table 8: *Unsupervised isiZulu AM training approach using a flat phone grammar.*

Data Set	Raw	Eval
Seed (1 hr)	31.8	33.59
+ 3 hr	31.44	33.62
+ 6 hr	32.45	37.62
+ 12 hr	35.91	44.54

Table 9: *Unsupervised isiZulu AM training approach using word LM.*

Data Set	Raw	Eval
Seed (1 hr)	29.11	30.65
+ 3 hr	29.0	31.84
+ 6 hr	30.36	35.68

- For SAE, the word LM gave expected performances, according to the performance metrics, which confirms the importance of a language model when using unsupervised acoustic model training.
- The amount of text data used to develop the LM has a slight effect on the performances: even with relatively small amounts of text, successful unsupervised training was achieved in SAE, though increasing performance with more data added to the training pool was observed. The drop in absolute performance may be related to the increasing OOV rates.
- Based on all the isiZulu results, the application of the unsupervised acoustic model training approach was unsuccessful – increasing amounts of automatically transcribed data produced poorer system accuracies. This is probably related to the high OOV rate of the isiZulu text corpus, which in turn results from the much larger vocabulary of a conjunctively written agglutinative language.
- Investigating the isiZulu results further showed: the correctness percentages for isiZulu increased, with increasing amounts of audio data but it was found that increasing the insertion penalty did not improve the accuracy values. This might suggest that the extremely high OOV rate severely limits the applicability of the unsupervised acoustic model training approach.

6. Future Work

For future work, it would be informative to investigate whether unsupervised acoustic model training for isiZulu can be made to work with a sufficiently large text source, but also to understand whether the approaches that do not require such a text source can be adapted to succeed in this context.

Another unknown for the isiZulu investigation is the effect of out-of-language words. English does not suffer from this phenomenon and in the majority of cases is the "invader" language in isiZulu. Out-of-Language words are particularly bothersome with respect to pronunciation modelling, and our phone-mapping approach is clearly a rough approximation in that case.

One possible approach to deal with the high isiZulu OOVs is to use a better syllabification approach. Our syllables did not yield any improvement over the word-based LM, but following the recent BABEL syllabification approach proposed by Davel

Table 10: *Unsupervised isiZulu AM training approach using syllable LM.*

Data Set	Raw	Eval
Seed (1 hr)	30.28	30.7
+ 3 hr	30.3	30.14
+ 6 hr	31.18	33.13
+ 12 hr	34.61	38.62

Table 11: *Unsupervised isiZulu AM training approach using word LM but starting with three hours of seed data.*

Data Set	Raw	Eval
Seed (3 hr)	25.44	29.47
+ 3 hr	26.29	30.7
+ 6 hr	27.66	32.68

et. al. [22] may help to achieve successful automatic isiZulu harvesting.

7. References

[1] T. J. Hazen, "Automatic alignment and error correction of human generated transcripts for long speech recordings," in *Proceedings of INTERSPEECH*. Pittsburgh, Pennsylvania, USA: ISCA, September 2006, pp. 1606–1609.

[2] L. Lamel, J. L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech & Language*, vol. 16, no. 1, pp. 115–129, 2002.

[3] C. Gollan, S. Hahn, R. Schlüter, and H. Ney, "An improved method for unsupervised training of LVCSR systems." in *Proceedings of INTERSPEECH*, Antwerp, Belgium, August 2007, pp. 2101–2104.

[4] L. Lamel, J.-L. Gauvain, and G. Adda, "Unsupervised acoustic model training," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. Norwich, UK: IEEE, May 2002, pp. I-877–I-880.

[5] S. Novotney, R. Schwartz, and J. Ma, "Unsupervised acoustic and language model training with small amounts of labelled data," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Taipei, Taiwan: IEEE, May 2009, pp. 4297–4300.

[6] L. Nguyen and B. Xiang, "Light supervision in acoustic model training," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. Montreal, Quebec, Canada: IEEE, May 2004, pp. I-185–I-188.

[7] H. Chan and P. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. Montreal, Quebec, Canada: IEEE, May 2004, pp. I-737–I-740.

[8] M. Gales, P. Woodland, H. Y. Chan, D. Mrva, R. Sinha, and S. E. Tranter, "Progress in the CU-HTK broadcast news transcription system," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1513–1525, 2006.

[9] M. H. Davel, C. van Heerden, N. Kleynhans, and E. Barnard, "Efficient harvesting of Internet audio for resource-scarce ASR," in *Proceedings of INTERSPEECH*. Florence, Italy: ISCA, August 2011, pp. 3153–3156.

[10] H. Gelas, L. Besacier, and F. Pellegrino, "Developments of Swahili resources for an automatic speech recognition system," in *SLTU-Workshop on Spoken Language Technologies for Under-Resourced Languages*, Cape Town, South Africa, May 2012.

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

[12] M. Davel and O. Martirosian, "Pronunciation dictionary development in resource-scarce environments," in *Proceedings of INTERSPEECH*, Brighton, United Kingdom, September 2009, pp. 2851–2854.

[13] J. Novak, D. Yang, N. Minematsu, and K. Hirose, "Initial and evaluations of an open source WFST-based phoneticizer," *The University of Tokyo, Tokyo Institute of Technology*.

[14] M. Davel and E. Barnard, "Pronunciation prediction with Default&Refine," *Computer Speech & Language*, vol. 22, no. 4, pp. 374–393, 2008.

[15] B.-J. Hsu and J. Glass, "Iterative language model estimation: efficient data structure & algorithms," in *Proceedings of INTERSPEECH*, vol. 8, Brisbane, Australia, September 2008, pp. 1–4.

[16] N. Kleynhans, R. Molapo, and F. De Wet, "Acoustic model optimisation for a call routing system," in *Proceedings of the Annual Symposium of the Pattern Recognition Association of South Africa*. Pretoria, South Africa: PRASA, November 2012, pp. 165–172.

[17] D. Kim, G. Evermann, T. Hain, D. Mrva, S. Tranter, L. Wang, and P. Woodland, "Recent advances in broadcast news transcription," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*. St. Thomas, U.S. Virgin Island: IEEE, November 2003, pp. 105–110.

[18] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1994, pp. 307–312.

[19] H. Kamper, F. de Wet, T. Hain, and T. Niesler, "Resource development and experiments in automatic sa broadcast news transcription," in *SLTU-Workshop on Spoken Language Technologies for Under-Resourced Languages*, Cape Town, South Africa, May 2012, pp. 102–106.

[20] R. Eiselen and M. Puttkammer, "Developing text resources for ten south african languages," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014.

[21] M. J. F. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low resource languages: Babel project research at cued," *Spoken Language Technologies for Under-Resourced Languages*, 2014.

[22] M. Davel, E. Barnard, C. van Heerden, W. Hartmann, D. Karakos, R. Schwartz, and S. Tsakalidis, "Exploring minimal pronunciation modeling for low resource languages," in *Accepted to Interspeech 2015*, 2015.