# Semi-Supervised Training for Lecture Transcription in Resource-Scarce Environments

Pieter de Villiers, Etienne Barnard, Charl J. van Heerden, Petri Jooste

Multilingual Speech Technologies Group
North-West University
Vanderbijlpark 1900, South Africa
Email: {pieterdevill, etienne.barnard, cvheerden, petri.jooste}@gmail.com

***Abstract*—We present a study where standard semi-supervised training methods are applied in a resource-scarce environment to build lecture transcription systems. Experiments are conducted on two different corpora which one can expect to be available in resource-scarce environments. These include 1) speaker- and domain-specific data where a single South African English lecturer presents the "Operating Systems" course, and 2) Afrikaans speaker-independent and domain non-specific data collected from *science* and *law* courses. Different amounts of acoustic and language model data are used for training the respective models. We find that lecture transcription systems in resource-scarce environments can benefit substantially from semi-supervised training methods. We also describe a small, new corpus of spoken lectures which is freely available in the public domain.**

***Index Terms*—Lecture Transcription, Kaldi, semi-supervised, Language Model, Resource-scarce.**

## I. Introduction

The availability of lecture transcriptions (LT) is known to be very rewarding in educational environments. LTs are useful to students as supplementary study material, especially for those students with disabilities which either prohibits or negatively influences their ability to create their own notes during a lecture. Given the large amounts of people with disabilities, it is clear that there is a growing need for automated LT systems.

Disabilities are surprisingly prevalent: there are approximately 2.35 million children (aged 6–21) in the UK who were reported to have disabilities [1], while the *National Institute on Disability and Rehabilitation Research* estimates that 15–20% of randomly selected people can have impairments considered as disabilities [2].

In an educational environment, students with learning disabilities often need more time to process learning material, while students with physical disabilities might struggle to take class notes themselves. This is a significant problem, as class notes have been identified as one of the most requested supplemental learning aids by students with disabilities [3]. In a study conducted by Ranchal et. al [3], a 10.5% increase in student grades was observed, after these students had been provided with multimedia class notes.

Class notes for students with disabilities are typically produced by student volunteers, as professional stenographers are too costly for everyday deployment [4], [5]. This task can become quite challenging for the students, as it is both time consuming and requires significant mental effort while also paying full attention to the lecturer [3]. College students were, for example, only able to capture ~40% of the information presented in a lecture [3]. Another study found that even with 2 volunteers, only 20–30% of a spoken lecture could be captured [5]. This problem is exacerbated when students have to take notes in a lecture presented in a language other than their mother tongue, which is often the case in the developing world. We hence believe that the potential benefit of LT systems may be even greater in the developing world, where lower literacy and a larger degree of multilingualism are more prevalent than in developed countries.

Implementing a LT system however, is a non-trivial process, with the development of the underlying automatic speech recognition (ASR) system being the main challenge. State-of-the art ASR systems typically require hundreds of hours of speech to train acoustic models (AMs) and millions of words to estimate reliable language models (LMs). Collating such resources are expensive and typically only exist in developed countries where there is an associated economic benefit. The resources necessary to build such systems in many languages of the developing world are either non-existent, or insufficient to reach the useful accuracy levels of resource-rich language LT systems. One feasible method of overcoming this challenge is by making use of semi-supervised training methods. This is typically an iterative process: an initial recognizer[1] will typically be used to create transcriptions of any untranscribed acoustic data. Well recognized pieces are then identified and extracted based on a confidence threshold and used (in combination with the original training data) to either adapt or retrain the AM. This can also be done in an iterative process [6], [7].

In this paper, we employ two lecture transcription systems built with resources expected to be available in resource-scarce environments. We show the benefit of semi-supervised training for such systems in a resource-scarce environment. We also investigate how different amounts of audio and domain specific LM training data influences the overall training process. These experiments are conducted on speaker- and domain specific

[1]This recognizer may either be a crude recognizer trained on a small amount of transcribed in-language data, or a well trained recognizer from another language with phone mappings used where necessary.

data (South African English), as well as speaker independent and domain non-specific data (Afrikaans).

In section II we provide some background on the development of ASR systems. Section III describes the corpora collected for AM development. Section IV describes development process followed during acoustic-, language- and pronunciation modeling. The experimental outcomes are listed in Section V with the conclusions drawn in Section VI.

## II. Background

Various lecture transcription systems have been implemented in well-resourced environments, where large amounts of transcribed audio and relevant language model data is available. An example of such a system is the MIT Spoken Lecture Processing project [8], where the developers had collected over 500 hours of recordings, of which more than 200 hours had been transcribed. Each lecturer had between 1–30 hours of speech which could be used for speaker adaptation, and the language models were trained on more than 6 million English words. This system achieved a Word Error Rate (WER) of 17% [9].

While lectures can easily be recorded, it is not feasible to generate comparable amounts of manual transcriptions to what was created for the MIT project, as it is both time consuming and expensive. Because of this, as well as the abundance of untranscribed data in multiple forms, unsupervised and semi-supervised training has become an attractive alternative [6], [7], [10].

Unsupervised training is known to significantly reduce WERs: Kemp et. al [11] found unsupervised training to decrease WERs from 32.1% to 20.6%, using as little as 30 minutes of transcribed and 50 hours of untranscribed data. Recognized word sequences were selected for adaptation/retraining if the confidence score was bigger than or equal to an empirically determined threshold of 0.5.

In another study, only 1.2 hours of transcribed data was used to train the initial recognizer [7]. This recognizer was then used in an iterative process (7 iterations) with a confidence threshold of 0.7 to decode 70.8 hours of untranscribed speech and subsequently retrain a new system. Using two evaluation sets (Broadcast News '96 and Broadcast News '98), they found a decrease in WER from 71.3% to 38.3% and from 65.5% to 29.3% respectively. They also reported a reduction in WER as the amount of initial transcribed data was increased.

## III. Corpus description

Two LT corpora were collected, with some of the lectures also manually transcribed: the Afrikaans Lecture Transcription corpus (ALT) and the English Operating Systems corpus (OS).[2]

[2]For information on the *OS* and *ALT* corpora, as well as free access to all the pronunciation dictionaries and lists, as used in the experiments reported in this paper, see https://sites.google.com/site/devilliers14lt/.

### A. Operating Systems corpus

The OS corpus [12] consists of a single male lecturer providing an OS course. While the lecturer's mother tongue is Afrikaans, he presents the course in English and speaks with a typical South African English accent. He has been presenting OS for several years and is thus able to arrive relatively "unprepared", recalling subject matter from memory. The lectures subsequently contain many false starts, corrections and hesitations.

There are 12 lectures in the OS corpus, ranging from 19–84 minutes per lecture; this amounts to ~12 hours of audio. Since smaller segments of data will result in faster alignment and decoding [13], the lectures were split into much smaller audio segments, ranging from less than one second, to ~40 seconds in duration. The audio segmentation was performed using Sox [14]; recordings were segmented based on a leading silence of 0.5 seconds at an audio threshold of 1%, and a trailing silence of 0.8 seconds at an audio threshold of 1%.

Six of the lectures were manually transcribed; 4 lectures as a training set, 1 lecture for development or tuning, and 1 lecture for evaluation. The remaining 6 lectures are untranscribed, and were used for semi-supervised training. Given our small collection of OS data, all experiments were performed using 6-fold cross-validation.

A summary of the OS corpus is shown in Table I, while Table II shows the data partitions used for each fold used in cross-validation.

TABLE I
*Segmented OS recordings with duration in minutes. The number of words for transcribed recordings are also shown.*

| ID | #Words | Dur.(min) |
|---|---|---|
| U1 | - | 33 |
| U2 | - | 8 |
| U3 | - | 53 |
| U4 | - | 47 |
| U5 | - | 31 |
| U6 | - | 11 |
| T1 | 5018 | 37 |
| T2 | 2446 | 17 |
| T3 | 2999 | 22 |
| T4 | 4358 | 28 |
| T5 | 6639 | 47 |
| T6 | 6766 | 47 |

TABLE II
*The OS data partitions for the 6 folds used for cross-validation. The IDs are listed in Table I.*

| Fold | Training set | Development set | Evaluation set |
|---|---|---|---|
| 1 | T1, T2, T3, T4 | T5 | T6 |
| 2 | T2, T3, T4, T5 | T6 | T1 |
| 3 | T3, T4, T5, T6 | T1 | T2 |
| 4 | T4, T5, T6, T1 | T2 | T3 |
| 5 | T5, T6, T1, T2 | T3 | T4 |
| 6 | T6, T1, T2, T3 | T4 | T5 |

### B. Afrikaans Lecture Transcription corpus

The ALT Corpus [15] consists of 20 hours of transcribed Afrikaans lectures from two general subject areas; *law* and

*science/chemistry*. Male and female lecturers account for 14 and 6 hours of lectures, respectively.

All audio and transcriptions were aligned using an AM trained on the Afrikaans NCHLT corpus [16], [15]. The aligned audio and transcriptions were then split into smaller segments using the Audacity [17] sound finder function. This resulted in audio segments ranging from less than one second, to about 30 seconds in duration.

A summary of all lecturers together with the subjects they presented, number of recordings and total duration in minutes, are listed in Table III.

All experiments were again conducted using n-fold cross-validation (n=5), given that the corpus is relatively small, yet contains subject matter from diverse disciplines.

The data partitions (from the different data sets) for each fold used for cross-validation are shown in Table IV.

TABLE III
*ALT Speaker Information with gender, subject type, number of recordings per lecturer and total duration in Minutes*

| ID | Gender | Subject | Recordings | Dur.(Segmented) |
|---|---|---|---|---|
| SP1 | male | sci | 3 | 102 |
| SP2 | male | sci | 2 | 78 |
| SP3 | male | law | 2 | 48 |
| SP4 | male | sci | 1 | 36 |
| SP5 | male | sci | 1 | 11 |
| SP6 | male | law | 3 | 89 |
| SP7 | male | law | 2 | 64 |
| SP8 | male | law | 1 | 46 |
| SP9 | male | law | 1 | 36 |
| SP10 | female | law | 3 | 92 |
| SP11 | female | law | 2 | 50 |
| SP12 | male | sci | 1 | 40 |
| SP13 | male | sci | 1 | 27 |
| SP14 | female | sci | 2 | 81 |
| SP15 | female | sci | 2 | 57 |
| SP16 | male | sci | 1 | 38 |
| SP17 | male | law | 1 | 25 |
| SP18 | male | sci | 2 | 72 |
| SP19 | male | law | 1 | 52 |
| SP20 | female | law | 1 | 38 |
| SP21 | female | sci | 1 | 19 |

## IV. Approach

Three experiments were conducted in order to determine how 1) the number of domain-specific audio transcriptions included in the LM, as well as 2) how the amount of initial acoustic training data affects the overall performance of a system trained using semi-supervised training methods.

These experiments can be summarized as follows:

- Include a *large* amount of audio transcriptions in the LM as well as a *large* amount of transcribed audio in the AM training data.
- Include a *large* amount of audio transcriptions in the LM but only a *limited* amount of transcribed audio in the AM training data.
- Include *no* audio transcriptions in the LM but a *large* amount of transcribed audio in the AM training data.

In Sections IV-A–IV-C we will describe our acoustic, language and pronunciation modelling approaches.

TABLE IV
*ALT data distribution for 5 folds of cross-validation. IDs are listed in Table III*

| | | |
|---|---|---|
| Fold 1 | Train | SP1, SP2, SP3, SP4, SP5, SP6, SP7, SP8, SP9 |
| | Dev | SP10, SP11, SP12, SP13 |
| | Eval | SP14, SP15, SP16, SP17 |
| | Untrans | SP18, SP19, SP20, SP21 |
| Fold 2 | Train | SP6, SP7, SP8, SP9, SP10, SP11, SP12, SP13 |
| | Dev | SP14, SP15, SP16, SP17 |
| | Eval | SP18, SP19, SP20, SP21 |
| | Untrans | SP1, SP2, SP3, SP4, SP5 |
| Fold 3 | Train | SP10, SP11, SP12, SP13, SP14, SP15, SP16, SP17 |
| | Dev | SP18, SP19, SP20, SP21 |
| | Eval | SP1, SP2, SP3, SP4, SP5 |
| | Untrans | SP6, SP7, SP8, SP9 |
| Fold 4 | Train | SP14, SP15, SP16, SP17, SP18, SP19, SP20, SP21 |
| | Dev | SP1, SP2, SP3, SP4, SP5 |
| | Eval | SP6, SP7, SP8, SP9 |
| | Untrans | SP10, SP11, SP12, SP13 |
| Fold 5 | Train | SP1, SP2, SP3, SP4, SP5, SP18, SP19, SP20, SP21 |
| | Dev | SP6, SP7, SP8, SP9 |
| | Eval | SP10, SP11, SP12, SP13 |
| | Untrans | SP14, SP15, SP16, SP17 |

### A. Acoustic Modeling

Fairly standard Kaldi [18] word recognition systems are trained using a recipe similar to the Kaldi Babel & WSJ recipes; our best results (optimized on a held out development set of similar size to the test set) are achieved with standard Gaussian Mixture Models (GMMs), using fMLLR speaker-specific transforms. The features employed are standard MFCCs with CMN per lecture (for the OS corpus) or per speaker (for the ALT corpus). Frames are spliced together, and LDA is used to reduce the dimensionality of the features to 40.

These experiments were conducted iteratively (up to a maximum of 3 iterations). After the initial low-accuracy models have been trained, the following steps were iteratively followed:

1) Decode any untranscribed data.
2) Extract word-based lattices.
3) Determine the best word-based confidence threshold based on the dev set.
4) Segment the original MFCC files based on word confidences.
5) Train new models using the new and initial data.
6) Repeat steps 1 – 5.

For experiments requiring only a limited set of training data, the OS model was trained using only a single lecture, while the ALT model was trained using two lectures (one from each of two individual speakers).

### B. Language modeling

Different LMs were employed during recognition and in combination with different acoustic models: these include LMs trained with and without audio transcriptions.

*1) OS corpus LM:* Four corpora were employed to train OS subject-specific LMs that were employed during decoding with the acoustic models described in Section III-A:

- *Lecturer* - manual transcriptions from the collected OS corpus (discussed in Section III-A).
- *OS Books* corpus - several English online books related to OS subjects.
- *Study guide* corpus - 2012 English study guides (collected from the North-West University Vaal Triangle campus), related to any Information Technology course.
- *Youtube* corpus - transcriptions uploaded to, or automatically generated by for example Google [19]. These include online tutorials on operating systems, as well as OS related subjects provided by Google talks.

Different LMs were trained for each fold of cross-validation. For each LM, the corresponding vocabulary was created by selecting all words present in the training set with a word frequency higher than 2. This heuristic proved to be useful for removing misspelled words.

Corpus-specific 3-gram LMs with Kneser Ney (KN) discounting were trained using SRILM [20]; the best interpolation weight for each corpus was then calculated on a held out dev set [6], [13].

Table V shows some statistics for LMs trained on the different corpora, evaluated on the development set of the first fold of cross-validation. The total number of words in each corpus, number of unigrams, number of 3-grams, perplexity (PPL), as well as % out-of-vocabulary (OOV) words are listed.

TABLE V
*LM statistics for different OS text corpora (evaluated on fold 1 dev set).*

| Corpus | #Words | Unigrams | #3-grams | PPL | %OOV |
|---|---|---|---|---|---|
| Lecturer | 17955 | 1549 | 1703 | 171.90 | 15.14 |
| OS Books | 1002827 | 20810 | 106702 | 258.01 | 3.60 |
| Studyguide | 157608 | 5988 | 18777 | 443.59 | 9.14 |
| Youtube | 277535 | 8875 | 27569 | 239.05 | 4.93 |

A total of 12 interpolated LMs were created; one for each fold of cross-validation, with LMs trained with and without transcription data. Some LM statistics for interpolated LMs are shown in Table VI.

TABLE VI
*Interpolated LM results, with and without audio transcriptions*

| Fold | Uni-grams | #3-grams | DEV PPL | DEV %OOV | EVAL PPL | EVAL %OOV |
|---|---|---|---|---|---|---|
| **With audio transcriptions** | | | | | | |
| 1 | 11380 | 143064 | 174.92 | 1.37 | 189.62 | 1.37 |
| 2 | 11380 | 143128 | 169.12 | 1.33 | 137.82 | 1.20 |
| 3 | 11383 | 143354 | 132.88 | 1.14 | 113.78 | 1.51 |
| 4 | 11387 | 143413 | 120.00 | 1.47 | 138.97 | 1.00 |
| 5 | 11388 | 143201 | 136.70 | 1.13 | 145.92 | 0.99 |
| 6 | 11385 | 143044 | 138.90 | 0.83 | 159.83 | 1.36 |
| **Without audio transcriptions** | | | | | | |
| 1 | 11356 | 142350 | 177.42 | 2.43 | 209.52 | 3.27 |
| 2 | 11356 | 142350 | 209.01 | 3.27 | 149.95 | 3.47 |
| 3 | 11356 | 142350 | 149.05 | 3.47 | 148.84 | 3.80 |
| 4 | 11356 | 142350 | 144.86 | 3.80 | 162.21 | 4.40 |
| 5 | 11356 | 142350 | 161.69 | 4.40 | 178.92 | 1.93 |
| 6 | 11356 | 142350 | 177.20 | 1.93 | 182.47 | 2.43 |

*2) ALT corpus LM:* Six corpora were used to train Afrikaans LMs that were employed during decoding with the acoustic models described in Section III-B. These corpora include the following:

- *Transcription* - manual audio transcriptions from the collected ALT corpus (discussed in Section III-B).
- *News*, *Wiki* and *Web* corpus - Three corpora collected from the Leipzig corpora collection site [21].
- *Study guide* (SG) corpus - all Afrikaans study guides related to law and natural sciences were collected from the NWU Vaal triangle and Potchefstroom campus.
- *Protea* corpus - used as a source for general proof-read text.

The vocabulary for these LMs were created by extracting the most frequent words from each source; the specific vocabulary cut-off was determined based on the reduction in OOV rate on the first fold's development set.

Table VII shows the results of the individually trained corpora, evaluated on the development set of the first fold of cross-validation.

TABLE VII
*LM statistics for different ALT text corpora (evaluated on fold 1 dev set).*

| Corpus | #Words | Unigrams | #3-grams | PPL | %OOV |
|---|---|---|---|---|---|
| Trans. | 74029 | 5210 | 6250 | 230.55 | 9.45 |
| News | 11228013 | 261330 | 1050546 | 652.47 | 3.25 |
| Web | 6934465 | 200027 | 652463 | 532.96 | 3.19 |
| Wiki | 2457950 | 130571 | 206984 | 911.44 | 5.20 |
| SG | 13839150 | 172578 | 1855316 | 619.67 | 2.02 |
| Protea | 6027884 | 136453 | 563266 | 646.28 | 4.35 |

A total of 10 interpolated LMs were trained; one for each fold of cross-validation, with LMs trained with and without audio transcriptions.

The results of these interpolated LMs are shown in Table VIII.

TABLE VIII
*Interpolated LM results, with and without audio transcriptions.*

| Fold | Uni-grams | #3-grams | DEV PPL | DEV %OOV | EVAL PPL | EVAL %OOV |
|---|---|---|---|---|---|---|
| **With audio transcriptions** | | | | | | |
| 1 | 22216 | 2808409 | 291.76 | 0.07 | 296.30 | 0.04 |
| 2 | 22216 | 2807869 | 326.26 | 0.04 | 325.89 | 0.05 |
| 3 | 22216 | 2808312 | 312.07 | 0.05 | 308.72 | 0.05 |
| 4 | 22216 | 2808212 | 305.23 | 0.05 | 267.12 | 0.05 |
| 5 | 22216 | 2808441 | 256.49 | 0.05 | 285.80 | 0.07 |
| **Without audio transcriptions** | | | | | | |
| 1 | 19463 | 2783506 | 302.46 | 5.44 | 310.83 | 5.76 |
| 2 | 19463 | 2783506 | 309.89 | 5.76 | 318.27 | 6.63 |
| 3 | 19463 | 2783506 | 313.81 | 6.63 | 363.42 | 4.94 |
| 4 | 19463 | 2783506 | 358.99 | 4.94 | 231.59 | 5.29 |
| 5 | 19463 | 2783506 | 228.92 | 5.29 | 305.61 | 5.44 |

From Tables VI and VIII, it is clear that the measurements on the evaluation sets are slightly worse than those on the development sets. This is mainly due to the process of LM interpolation which makes use of the development set to "fine tune" the models. As expected, models containing transcription data in the training sets perform better than models without.

### C. Pronunciation modeling

Pronunciation dictionaries were created (1) using a dictionary lookup for known words, (2) identifying foreign words with a dictionary lookup and (3) using the Default & Refine [22] algorithm to automatically predict pronunciations for the remaining words. Pronunciations of foreign and predicted words were then manually checked, while pronunciations of foreign words were mapped to their appropriate phones [23].

The OS and ALT pronunciation dictionaries contained 17370 and 22214 words respectively.

## V. Experiments

As mentioned in Section IV, three experiments were conducted using the OS and ALT models.

The results for these three experiments on both the OS and ALT models are shown in Tables IX to XI and Tables XII to XIV respectively. The results are averaged across all folds of cross-validation (6 folds for the OS data and 5 folds for the ALT data). The WERs together with their Standard Errors, best confidence threshold based on the dev set, as well as the total data extracted (for use as training material during the next iteration) are shown.

For each experiment, only 3 iterations of semi-supervised training was performed (it was empirically determined on the dev set that very little improvement occurs after 3 iterations. (Iteration 0 represents the initial models.)

TABLE IX
*OS average results over all 6 folds. Transcribed lectures are included in both the LM as well as the AM.*

| Iter | DEV WER | EVAL WER | Conf. Thres. | Words extracted | Minutes extracted |
|---|---|---|---|---|---|
| 0 | 37.33 ± 2.48 | 37.00 ± 2.22 | 0.81 | 24270 | 2:12:16 |
| 1 | 34.42 ± 1.88 | 34.38 ± 1.77 | 0.77 | 24941 | 2:19:15 |
| 2 | 34.35 ± 1.89 | 34.57 ± 1.70 | 0.69 | 25441 | 2:21:19 |
| 3 | 34.60 ± 2.16 | 34.97 ± 1.72 | - | - | - |

TABLE X
*OS average results over all 6 folds. Transcribed lectures are included in the LM, but only a limited amount of transcribed lectures are included in the AM.*

| Iter | DEV WER | EVAL WER | Conf. Thres. | Words extracted | Minutes extracted |
|---|---|---|---|---|---|
| 0 | 49.45 ± 2.93 | 49.80 ± 1.68 | 0.90 | 20516 | 1:49:32 |
| 1 | 40.27 ± 2.39 | 40.72 ± 1.35 | 0.81 | 24308 | 2:11:42 |
| 2 | 39.47 ± 1.98 | 39.42 ± 1.64 | 0.83 | 24432 | 2:12:09 |
| 3 | 39.35 ± 2.06 | 39.68 ± 1.47 | - | - | - |

TABLE XI
*OS average results over all 6 folds. While transcribed lectures are included in the AM, no transcribed lectures are included in the LM.*

| Iter | DEV WER | EVAL WER | Conf. Thres. | Words extracted | Minutes extracted |
|---|---|---|---|---|---|
| 0 | 38.88 ± 2.29 | 38.67 ± 2.06 | 0.79 | 24266 | 2:11:40 |
| 1 | 36.12 ± 1.73 | 36.18 ± 1.80 | 0.78 | 24655 | 2:17:00 |
| 2 | 35.87 ± 1.75 | 36.02 ± 1.70 | 0.79 | 24780 | 2:18:02 |
| 3 | 36.03 ± 1.69 | 36.08 ± 1.77 | - | - | - |

TABLE XII
*ALT average results over all 5 folds. Transcribed lectures are included in both the LM as well as the AM.*

| Iter | DEV WER | EVAL WER | Conf. Thres. | Words extracted | Minutes extracted |
|---|---|---|---|---|---|
| 0 | 50.70 ± 2.04 | 51.18 ± 2.09 | 0.78 | 20673 | 1:53:04 |
| 1 | 49.04 ± 1.83 | 49.56 ± 1.89 | 0.78 | 23511 | 2:08:48 |
| 2 | 49.14 ± 1.85 | 49.38 ± 2.06 | 0.76 | 24397 | 2:13:54 |
| 3 | 48.92 ± 1.91 | 49.56 ± 1.90 | - | - | - |

TABLE XIII
*ALT average results over all 5 folds. Transcribed lectures are included in the LM, but only a limited amount of transcribed lectures are included in the AM.*

| Iter | DEV WER | EVAL WER | Conf. Thres. | Words extracted | Minutes extracted |
|---|---|---|---|---|---|
| 0 | 84.44 ± 3.01 | 85.66 ± 2.47 | 0.51 | 12687 | 1:01:30 |
| 1 | 78.54 ± 3.98 | 79.20 ± 4.19 | 0.80 | 14803 | 1:18:45 |
| 2 | 75.04 ± 4.46 | 76.54 ± 5.00 | 0.97 | 15063 | 1:22:24 |
| 3 | 72.42 ± 4.05 | 74.46 ± 4.88 | - | - | - |

TABLE XIV
*ALT average results over all 5 folds. While transcribed lectures are included in the AM, no transcribed lectures are included in the LM.*

| Iter | DEV WER | EVAL WER | Conf. Thres. | Words extracted | Minutes extracted |
|---|---|---|---|---|---|
| 0 | 56.46 ± 2.00 | 57.08 ± 2.13 | 0.87 | 18827 | 1:41:05 |
| 1 | 55.22 ± 1.94 | 55.74 ± 1.93 | 0.85 | 22169 | 1:59:01 |
| 2 | 55.20 ± 1.95 | 55.74 ± 2.06 | 0.84 | 23223 | 2:05:28 |
| 3 | 55.20 ± 1.93 | 55.76 ± 2.12 | - | - | - |

## VI. Conclusion

Experiments were conducted on two South African lecture transcription corpora, investigating different variables relevant to semi-supervised training. It was found that there is little to no gain after 3 iterations of semi-supervised training, and that a confidence threshold of ~0.80 works well.

From Tables IX and X and Tables XII and XIII it is clear that an increase of ~99 minutes of acoustic training data in the OS model and ~410 minutes in the ALT model, resulted in a relative improvement of 11.87%, and 33.44% respectively.

A similar but less pronounced trend is observed when considering the amount of transcription data used in the LM (3.08% relative improvement in the OS model and 11.12% in the ALT model) (see Tables IX, XI, XII and XIV).

While the amount of transcribed acoustic training data has a larger impact on the accuracy of the recognizer than the amount of corresponding transcription data included in the language model's training data for the amount of data we experimented with, we expect the latter trend to become more pronounced as the amount of transcribed data increases. Future work should include exploring the impact of significantly increasing the amount of transcribed data available for acoustic and language modeling on recognition accuracy.

In conclusion, we have demonstrated the benefit of semi-supervised training for LT in resource-scarce environments on two new South African LT corpora.

## REFERENCES


[1] Anon., "Data accountability center: individuals with disabilities education act(idea) data," www.IDEAdata.org, 2011, accessed: 2013-10-25.

[2] K. Bain, S. H. Basson, and M. Wald, "Speech Recognition in University Classrooms : Liberated Learning Project," in *Proceedings of the fifth international ACM conference on Assistive technologies - Assets '02*. New York, New York, USA: ACM Press, Jul. 2002, pp. 192–196.

[3] R. Ranchal, T. Taber-Doughty, Y. Guo, K. Bain, H. Martin, J. P. Robinson, and B. S. Duerstock, "Using speech recognition for real-time captioning and lecture transcription in the classroom," *IEEE Transactions on Learning Technologies*, vol. 99, pp. 1–14, 2013.

[4] T. Kawahara, "Automatic transcription of parliamentary meetings and classroom lectures-a sustainable approach and real system evaluations," in *Proceedings of the 7th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Tainan, Taiwan, December 2010, pp. 1–6.

[5] T. Kawahara, N. Katsumaru, Y. Akita, and S. Mori, "Classroom note-taking system for hearing impaired students using automatic speech recognition adapted to lectures." in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech)*, Makuhari, Japan, September 2010, pp. 626–629.

[6] J. Lööf, C. Gollan, and H. Ney, "Cross-language bootstrapping for unsupervised acoustic model training: rapid development of a polish speech recognition system." in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech)*, Brighton, United Kingdom, September 2009, pp. 88–91.

[7] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Madonna di Campiglio, Italy, December 2001, pp. 307–310.

[8] J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent Progress in the MIT Spoken Lecture Processing Project," in *Interspeech 2007 (8th Annual Conference of the International Speech Communication Association)*. Antwerp, Belgium: ISCA, Aug. 2007, pp. 2553–2556.

[9] J. R. Glass, T. J. Hazen, D. S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent progress in the mit spoken lecture processing project." in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech)*, Antwerp, Belgium, August 2007, pp. 2553–2556.

[10] T. Kawahara, "Automatic transcription of parliamentary meetings and classroom lectures - A sustainable approach and real system evaluations -," in *7th International Symposium on Chinese Spoken Language Processing (ISCSLP*, Nov. 2010, pp. 1–6.

[11] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: recent experiments." in *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech)*, Budapest, Hungary, September 1999, pp. 2725–2728.

[12] P. T. de Villiers, "Lecture transcription systems in resource-scarce environments," Master's thesis, North-West Univerity Vaal triangle campus, May 2014.

[13] C. J. van Heerden, P. de Villiers, E. Barnard, and M. H. Davel, "Processing spoken lectures in resource-scarce environments," in *Proceedings of the 22nd Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Vanderbijlpark, South Africa, November 2011, pp. 138–143.

[14] Anon., "Sox - sound exchange," http://sox.sourceforge.net/, 2013, accessed: 2013-10-31.

[15] P. de Villiers, P. Jooste, C. J. van Heerden, and E. Barnard, "Towards lecture transcription in resource-scarce environments," in *Proceedings of the 23rd Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Pretoria, South Africa, November 2012, pp. 138–143.

[16] E. Barnard, M. H. Davel, C. van Heerden, F. de Wet, and J. Badenhorst, "The NCHLT Speech Corpus of the South African languages," in *Proc. SLTU*, St Petersburg, Russia, May 2014.

[17] Anon., "Audacity," http://audacity.sourceforge.net/, 2014, accessed: 2014-03-20.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December 2011, iEEE Catalog No.: CFP11SRW-USB.

[19] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Olomouc, Czech Republic, December 2013, pp. 368–373.

[20] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proceedings of ICSLP*, vol. 2, Denver, USA, September 2002, pp. 901–904.

[21] C. Biemann, G. Heyer, U. Quasthoff, and M. Richter, "The leipzig corpora collection-monolingual corpora of standard size," *Proceedings of Corpus Linguistic 2007*, 2007.

[22] M. Davel and E. Barnard, "Pronunciation predication with Default&Refine," *Computer Speech and Language*, vol. 22, pp. 374–393, Oct. 2008.

[23] C. J. van Heerden, P. de Villiers, E. Barnard, and M. H. Davel, "Processing Spoken Lectures in Resource-Scarce Environments," in *PRASA2011 - Proceedings of the 22nd Annual Symposium of the Pattern Recognition Association of South Africa*, P. Robinson and A. Nel, Eds., Vanderbijlpark, South Africa, 2011, pp. 138–143.