Towards Lecture Transcription in Resource-Scarce Environments

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

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

***Abstract*—We present progress towards automated Lecture Transcription (LT) in resource scarce environments. Our development has focused on the transcription of lectures in Afrikaans from two faculties at North-West University. A bootstrapping procedure is followed to filter and select well-aligned segments of speech. These segments are then used to train acoustic models. Initial work towards language modeling for LT in a resource-scarce environment is also presented; manual lecture transcriptions are combined with text mined from other sources such as study guides to train language models. Interpolation results indicate that study guides are a useful resource for language modeling, whereas general text (obtained from a publisher of Afrikaans books) is less useful in this context. Our findings are confirmed by the reduced word error rates (WERs) obtained from our off-line speech-recognition system for Lecture Transcription.**

***Index Terms*—Lecture Transcription, Afrikaans, Kaldi, Dynamic Programming, Language Model, Resource-scarce.**

I. Introduction

The availability of lecture transcriptions is understood to be very rewarding – most obviously for students with hearing disabilities, but also for the larger student population. Students with hearing disabilities use these transcriptions as a supportive learning medium, while students without such disabilities use them to better understand the lecturer or to supplement their class notes [1]. The multilingual environment of countries such as South Africa offers additional motivation for the development of lecture transcriptions, since students often attend lectures in languages other than their first language, and can therefore obtain significant benefit from transcriptions (either in real-time or off-line).

Hence, Kawahara et al. [2] report that some universities use student volunteers to create notes of classes, since professional stenographers are too costly. However, real-time transcription of lectures is infeasible for humans, and it was found that with 2 volunteers only 20-30% of the spoken lecture utterances could be transcribed in real-time. Another drawback is that these volunteers have to be familiar with the field of the lecture to be able to recognize domain-specific technical words. As a consequence, automated systems for lecture transcription, even with limited accuracy and topic coverage, hold great promise in multilingual universities.

Previous work on lecture transcription for Afrikaans [3] focused on different approaches to alignment, in order to harvest enough data from approximately transcribed lectures to retrain acoustic models using both a well trained target-language (Afrikaans) acoustic model as well as an acoustic model from another language. It was found that the target-language acoustic model performs significantly better for this task.

Given these results, as well as the availability of audio data collection applications such as Woefzela [4] and smart phones, we do not consider obtaining sufficient target-language audio for acoustic modeling as big an obstacle as it was in the recent past,[1] although the optimal approach to combining general and speaker-specific audio data in this context remains an interesting topic for investigation.

The current main challenge therefore with lecture transcription systems in resource-scarce environments is language modeling: lecturers tend to use domain specific words, spontaneous speech containing many false starts, hesitations, filled pauses, non-lexical artifacts such as coughs and laughs, and many other phenomena present in daily human communication [1], [5]. All these affect the accuracy of the speech recognition system. This is even more challenging in resource-scarce environments where very little text data is typically available for accurately modeling these artifacts with language models.

In this paper, we present results from our Afrikaans Lecture Transcription system. Our acoustic modeling approach is described in Section III-C2, and is similar to the approach described in [3], relying heavily on the Dynamic Programming-based audio harvesting procedure described in [6]. We employ a significantly expanded Afrikaans Lecture Transcription (ALT) corpus compared to that in [3], however, enabling us to work with a larger corpus, experiment more thoroughly with speaker-adaptive training, language modeling and also perform actual lecture transcription.

In Section III-F2, we present initial promising results when interpolating language models trained on text resources one can typically expect to exist even in resource-scarce environments: a small amount of transcribed lecture text and a much larger collection of text obtained from study guides. The effect

[1]Woefzela is a freely available Android application and can be used as a medium for collecting data in typical developing-world environments. It provides the user with a reliable and cost effective way of collecting target-language data even in remote locations.

of speaker adaptive training is investigated in Section III-E.

II. BACKGROUND

Various lecture transcription systems, such as the MIT Spoken Lecture Processing project [5], have been implemented in well-resourced environments. For that system, the developers had collected over 500 hours of recordings, of which over 200 hours had been transcribed. For the purposes of speaker adaptation, that corpus contained between 1 to 30 hours of speech per speaker, and the language models were trained on more than 6 million English words.

According to Munteanu et al. [7], current lecture transcription systems obtain word error rates (WER) between 40% and 45% whilst a minimum WER of 25% is acceptable by users. Even though recognition accuracies as high as 98% have been reported in certain Automatic Speech Recognition (ASR) systems, such high accuracies invariably require extremely favorable conditions, such as reading selected materials (from a limited context) aloud [1].

Glass et al. [5] found a greater improvement in WER from acoustic modeling than language modeling. However, they found that performing acoustic modeling on four 50 minute lectures from a single lecturer, while also performing language model adaptation using two related textbooks and 40 related lectures, still resulted in a high WER (30.7%). Using 29 hours of previous lectures for acoustic modeling decreased the WER noticeably(17%). Similar results where found by Trancoso et al. [8].

Unsupervised training is currently also receiving significant attention. Here, term discovery algorithms are used to identify words or phrases of different speakers and genders by identifying repetitions in the data. Jansen and Church [9] demonstrated that unsupervised training of acoustic models is possible with strong speaker independent properties.

III. APPROACH

Throughout our experiments we made use of two speech corpora: the NCHLT corpus [4] and the Afrikaans Lecture Transcription corpus, which was developed to support the current research.

A. NCHLT corpus

The NCHLT corpus consists of speech from 206 Afrikaans speakers (approximately equal numbers of males and females), with approximately 500 3-5 word utterances of read speech per speaker, recorded in a controlled environment. This amounts to approximately 100 hours of speech data. The vocabulary of this corpus consists of 9375 distinct words, drawn from a variety of subjects, as would be appropriate (for example) for a Web-search application.

B. Afrikaans Lecture Transcription corpus

The Afrikaans Lecture Transcription (ALT) Corpus consists of 20 hours of Afrikaans lecture data from two broad subject areas; law and science/chemistry. Male lecturers account for 14 hours of speaker data and females 6 hours. All audio data has been manually segmented into 5 minute segments, mainly to increase the speed of the alignment and decoding [3].

A single first-language Afrikaans speaker produced orthographic transcriptions of the ALT corpus; the transcriber was given the following instructions:

- Transcribe exactly what was said (do not correct for grammar, hesitations, etc)
- Use punctuation (,.?!) only to indicate sentence structure (no quotation marks or brackets)
- Write out numbers in words instead of using digits 0-9
- Mark foreign words with #

All speakers are listed in Table I with their associated subjects, gender and amount of training and testing data in minutes. The test set consists of one lecture from each of those lecturers who has multiple lectures in the ALT corpus.

TABLE I
ALT SPEAKER INFORMATION WITH TRAINING AND TESTING DATA IN MINUTES

SPKR ID	Gender	Subject	Train	Test	Total
m001	male	sci	17	0	17
m002	male	sci	42	37	79
m003	male	sci	84	37.5	121.5
m004	male	sci	31	0	31
m005	male	sci	44	0	44
m006	male	sci	46	37	83
m007	male	sci	43	0	43
m008	male	sci	37	0	37
m009	male	law	26	23	49
m010	male	law	36	0	36
m011	male	law	35	35.5	70.5
m012	male	law	62.5	37.5	100
m013	male	law	57.5	0	57.5
m014	male	law	47	0	47
m015	male	law	27	0	27
f001	female	sci	39.5	23	62.5
f002	female	sci	46.5	43	89.5
f003	female	sci	25	0	25
f004	female	law	32.5	30.5	63
f005	female	law	61.5	36	97.5
f006	female	law	40.5	0	40.5

C. Baseline systems for alignment

Four baseline systems were created for the purposes of alignment and subsequent harvesting of well-transcribed portions of the ALT corpus. This was done by employing the iterative DP scoring and filtering technique described in [6]. Before we describe the four systems in more detail, we will first elaborate on the experimental setup followed for the alignment systems.

1) Pronunciation modeling: Pronunciation dictionaries were created for all systems by (1) using a dictionary lookup for known Afrikaans words (443 words), (2) identifying English words with a dictionary lookup (840 words) and (3) using the Default & Refine [10] algorithm to automatically generate pronunciations for the remaining 6735 words.

English words occur fairly frequently in the ALT corpus; they were automatically identified by a dictionary lookup and the pronunciation mapped to similar Afrikaans phones was the same as in [3]. These mappings are shown in II. All names and

foreign words (which were marked with # by the transcriber) were then manually verified.

TABLE II
ENGLISH TO AFRIKAANS PHONE MAPPINGS

Eng	Afr	Eng	Afr
3:	@	Q	O
e@	E	r\	r
ai	a i	tS	t S
au	a u	u:	u
d_0Z	d Z	U	u
i:	i	T	f
O:	O	D	v
Oi	O i		

2) Acoustic modeling: The acoustic models for alignment were trained on 39 dimensional Mel frequency cepstral coefficients (13 static, 13 deltas and 13 double deltas). Offline cepstral mean and variance normalization was applied per speaker (that is, the same normalization constants were applied to all the speech from one speaker, and these constants were computed so that all speakers have the same cepstral means and variances after normalization). The hidden Markov models (HMMs), trained with HTK [11], were standard 3-state left to right tied-state triphone models, with 8 mixtures per state and semi-tied transforms. A garbage model [6] was then trained and combined with the initial model.

The following acoustic models were trained:

- **NCHLT baseline**. An acoustic model was trained on the NCHLT corpus described in Section III-A. This model was trained without a garbage model.
- **ALT (5-minute segments)**. We trained acoustic models using the entire manually segmented ALT corpus. Segments were approximately 5 minutes in duration. This acoustic model resulted in a phone accuracy of 45.14%. Based on our earlier experience with this corpus, this was good starting accuracy for a baseline system; we nevertheless decided to make use of DP scoring to automatically further segment the ALT data into smaller – but more reliable – segments, that could be used for further training.
- **DP filtered ALT**. The ALT corpus was automatically segmented into 10 second or smaller chunks as done by [5]. Our process employed the dynamic-programming phone string alignment procedure used by [3] with a flat phone matrix. As described in [3], this approach compares the result of a forced alignment with that of a free decode using a variable cost matrix, identifies the accurately transcribed sections of audio and use these results to segment the audio as well as the transcriptions. Using this technique we segmented the 5 minute ALT data into small chunks of accurately transcribed data using the NCHLT model. These well-aligned portions were then used to train a new improved ALT model.
- **NCHLT MAP**. The NCHLT model was then also MAP adapted using the entire ALT training set and used to automatically segment the ALT data into small chunks of accurately transcribed data using the dynamic-programming technique.

D. Alignment results

The phone accuracies of these 4 baseline systems, tested on the same ALT data are shown in Table III. Here the reference phone strings were generated by using the pronunciations as described in Section III-C1 since it was infeasible to obtain manual phone transcripts. The improvements in various measures of alignment accuracy (see [6] for motivations) after model refinement are shown in Table IV[2].

TABLE III
PHONE-RECOGNITION ACCURACIES OF BASELINE SYSTEMS TESTED ON ALT

NCHLT	LT(5 min)	LT(DP scoring)	NCHLT(MAP all)
19.28%	45.14%	49.70%	16.50%

As seen in Table III, domain-specific training data is very beneficial for the development of a baseline system. A further significant increase in phone accuracy (nearly 5%) is achieved by segmenting the ALT(5min) training data using the dynamic-programming technique.

TABLE IV
MEASURES OF ALIGNMENT ACCURACY ACHIEVED AFTER MODEL REFINEMENT ON THE TEST SET

Model	Avg DP Score	Log P	Time
NCHLT	-0.176	-52.10	4:05/5:39
NCHLT-MAP/all (ALT)	-0.217	-50.82	3:53/5:39
NCHLT MAP/spk	-0.202	-51.03	3:35/5:39
ALT	0.114	-45.60	3:14/5:39

E. Speaker Adaptation

Speaker adaptation was performed on multiple speakers for which we had data from more than one lecture. One or more lectures were used for speaker adaptation (Train column, Table I), and one lecture was held out for testing purposes (Test column, Table I). The NCHLT corpus acoustic model was also adapted to these speakers to see how important the use of speaker-specific data is to the overall system used for alignment. Table V summarizes the phone accuracies for different speakers on ALT without MAP adaptation, ALT with MAP adaptation, NCHLT without MAP adaptation and the NCHLT model with MAP adaptation. These results were obtained using the same techniques as in [8] where 3 iterations of speaker adaptation is performed using the same adaptation data.

We see that some speakers achieve only small gains in phone accuracy, and reduced accuracies after adaptation are even seen in many cases. These disappointing results are probably a consequence of the small amount of adaptation data

[2]It would have been preferable to make these measurements on a held out development set, but because of data scarcity and since these measures do not influence our decision on which model to use for the final data segmentation, we measured the model refinement on the test set

TABLE V
PHONE ACCURACY PER SPEAKER WITH AND WITHOUT MAP ADAPTATION, USING DIFFERENT BASELINE ACOUSTIC MODELS

SPKR ID	ALT	ALT + MAP	NCHLT	NCHLT + MAP
m002	59.69	62.12	27.48	29.01
m003	66.82	67.42	33.95	38.03
m006	48.84	49.59	12.38	12.20
m009	50.73	52.21	19.48	18.66
m011	59.39	61.92	19.97	18.72
m012	55.29	49.02	18.60	15.94
f001	39.24	39.43	16.99	15.45
f002	39.91	39.04	13.53	13.77
f004	40.93	34.97	17.84	15.68
f005	28.14	28.14	12.29	8.50
Avg	48.9	48.39	19.25	18.6

available to us – thus, the risk of overtraining is significant, and MAP adaptation is not able to compensate for the differences in recording conditions between the two corpora.

F. Baseline systems for lecture transcription

The process described in Section III-C is useful in a resource-scarce environment where only a few hours of lecture transcription data is available. The resulting well-aligned portions of our ALT corpus were used to train state of the art acoustic models for offline lecture transcription.

1) Acoustic modeling: The Kaldi toolkit [12] was used to train our best acoustic models which were used for transcribing lectures. Standard MFCCs with cepstral mean normalization were again used, with LDA, MLLT, speaker adaptive training (MLLR) and boosted mmi.

2) Language modeling: The most basic language model for this task is a simple word trigram model, built from the transcribed lectures. The transcriptions were separated into the two groups (law and natural sciences), and each sub-corpus was used for a specialized language model (in addition to the basic language model covering both topics). Because of the small corpus size of these transcriptions, we wanted to investigate whether recognition accuracy could be improved by using larger language models from a more general source of Afrikaans text. The Puk-Protea-Boekhuis corpus was used for this purpose. It contains Afrikaans text from published works and contains substantial quantities of proofread material on various topics. (In addition to prose and instruction documents, it also includes poetry.) In an effort to improve on the anticipated out-of-vocabulary words in the university lectures, a general set of study guides of the North West University was obtained. Since the available lectures chosen to be transcribed were from two different faculties, law and natural sciences, the study guides of these two faculties were used as corpus for the experiments to be described here.

The following three groups of corpora were therefore used: 1) Transcriptions of lectures, 2) University study guides, 3) General text. The details of these corpora are as follows:

- *Transcriptions of lectures in law (1A-tr-law) and natural sciences (1B-tr-sci).* Only the transcriptions of the training data was used so that the test data for measuring the recognition performance would not be in the language models. These were the smallest corpora 60K and 55K words respectively.
- *Study guides for subjects in law (2A-sg-law) and natural sciences (2B-sg-sci).* Only the Afrikaans versions of the study guides were used. For some subjects the study guides were bilingual combining Afrikaans and English in the same document, but these were not used. After text normalisation, these corpora were 1.4 million and 2 million words respectively.
- *The Puk-Protea-Boekhuis (protea) corpus* was used as a source for general proof-read text. After normalisation described below this corpus consisted of 6 million words.

Several steps were taken when pre-processing these text corpora for building language models. These issues are addressed in the discussion on text normalisation below. Initial experiments with 2-, 3-, 4-, 5-, and 6-grams showed lowest perplexity in most cases with 3-grams. In a few cases 4-grams were marginally better, but the benefits were never sufficiently large to justify the cost of the larger language models. All language models reported below are therefore trigram word models, which were built using the Modified Kneser-Ney smoothing algorithm [13]. In all models, the markers for beginning and end of sentences are included as tokens.

To get a basic measure of perplexity and out-of-vocabulary rate for each language model, the text from the lecture transcription test data was used for testing. The test set was also divided into *law* and *natural sciences* transcriptions; language-modeling results when testing on these two sub-corpora are reported in Table VI.

TABLE VI
PPL (PERPLEXITY) AND OOV (OUT-OF-VOCABULARY) RATE FOR BASELINE LANGUAGE MODELS.

Corpus	#2-grams	#3-grams	Test set	PPL	OOV rate
1A-tr-law	28725	5524	law	171.94	7.92%
1B-tr-sci	25285	5169	law	160.08	16.21%
2A-sg-law	255162	117280	law	404.96	7.01%
2B-sg-sci	421448	231741	law	647.88	9.33%
protea	1261554	446814	law	443.23	6.20%
1A-tr-law	28725	5524	sci	174.36	15.39%
1B-tr-sci	25285	5169	sci	151.53	7.54%
2A-sg-law	255162	117280	sci	664.32	15.22%
2B-sg-sci	421448	231741	sci	673.49	6.84%
protea	1261554	446814	sci	498.24	8.86%

As expected, the in-domain transcriptions provide the best match for the respective test sets, with relatively low perplexities and OOV rates. Interestingly, the cross-domain perplexities are comparable to the in-domain perplexities, with *tr-sci* actually achieving a somewhat lower perplexity on the *law* test set (although at a substantially higher OOV rate). Also, the OOV rates from the study guides and *Protea* corpus are encouragingly low, suggesting that corpus combination may be a profitable strategy. However, the disparate sizes of the corpora indicates that combination through weighted interpolation (rather than pooled resources) should be the strategy of choice; we therefore investigated the characteristics

of various interpolated language models involving these five sub-corpora.

3) Interpolated language models: To investigate the potential benefits of language-model interpolation, we created test sets from each of the five sub-corpora mentioned above. Modified Kneser-Ney smoothing was employed to estimate language models based on the training sets extracted from the same corpora, and the SRILM toolkit [14] was then used to find the optimal interpolation weights for combining these language models.

Because these various corpora have widely different characteristics, it is not meaningful to compare the perplexities and OOV rates across different test sets. We therefore focus on the interpolation weights that yield the lowest perplexity for each test set, as shown in Table VII.

TABLE VII
INTERPOLATION WEIGHTS THAT MINIMIZE THE PERPLEXITIES ON FIVE SUB-CORPORA.

Test Corpus	Weight: 1A-tr-law	Weight: 1B-tr-sci	Weight: 2A-sg-law	Weight: 2B-sg-sci	Weight: Protea
1A-tr-law	0.415	0.137	0.266	0.016	0.165
1B-tr-sci	0.074	0.606	0.004	0.170	0.145
2A-sg-law	0.001	0.000	0.933	0.042	0.024
2B-sg-sci	0.000	0.000	0.024	0.960	0.015
Protea	0.003	0.002	0.008	0.007	0.979

As expected, the diagonal entries in Table VII dominate; that is, the training set of each corpus makes the largest contribution to the lowest-perplexity language models of the corresponding test set. However, for the two LT test sets (the first two rows of the table), the other corpora also contribute substantially to the optimal language models. Interestingly, it is the study guide from the same domain as the test set which makes the largest contribution in each case; this is evidence that the non-transcription corpora are contributing to these language models in a predictable way.

4) Recognition results with interpolated corpora: Based on the analysis in Section III-F2 and Section III-F3, it was decided to use the transcriptions of lectures (LT) and University study guides (SG) to train language models to be used for off-line lecture transcription. Our goal here is to investigate whether improved recognition performance can be achieved with language-model interpolation, and to simplify the presentation we focus on the within-topic interpolation of the study guides and transcriptions. That is, we build two sets of language models, one for the *law* domain and the other for the *sciences* domain. In each set, we investigate the effect of ranging the interpolation weight between 0.0 (at which value the study guides dominate) to 1.0 (where the model is dominated by the transcriptions). For consistency, we report all results at a LM weight of 14, which is a reasonable value for our configuration, but not optimized for a particular language model.

As Figures 1, 2 and 3 show, we find in all cases, and for both the *law* and *sciences* test sets, that optimal performance is achieved at an interpolation value somewhere between the extremes, thus showing that language-model interpolation

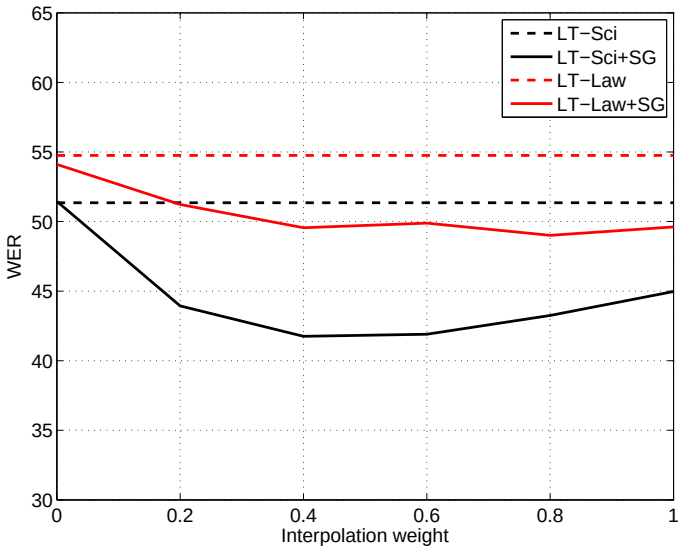

Fig. 1. WER for off-line lecture transcription when trained on *sci* and *law* sources respectively and evaluated on the combined *sci* and *law* LT test set. The dotted lines correspond to language models trained only on the LT training data, and the solid lines represent interpolated results, with the interpolation weight on the horizontal axis.

indeed is beneficial in all cases. As could be expected, the in-domain language models perform best on both test sets; these differences in word error rates are quite large, confirming the importance of language modeling for this task.

Note that our approach to interpolation requires a fixed vocabulary for all settings of the interpolation weight. Therefore, all words from both training sets are included in all interpolated models, albeit with only the unigram back-off probabilities in some cases. In Figures 1, 2 and 3 we can see that even these unigrams make a useful contribution to recognition accuracy, since the WERs with only the lecture-transcription language models (dotted lines in Figs.1, 2 and 3) are notably higher than the corresponding interpolated models (right-most points of the solid lines).

IV. CONCLUSION

Whereas the use of target-language acoustic data has previously shown to be beneficial [3], we have additionally demonstrated that domain-specific training data significantly contributes to the accuracy of lecture transcription systems. This is true even if the amount of available in-domain data is severely limited. However, under these constraints, the additional value of speaker adaptation is minimal.

We have also shown that additional target-language text, such as study guides, can lead to a substantial reduction in word error rates. Since such sources are likely to be available in the type of educational environment which is expected to represent the most important use case of this technology, this is a practically important result.

The error rates that we have achieved are still somewhat higher than those that are considered usable in lecture-transcription applications [7]; hence, the need for further

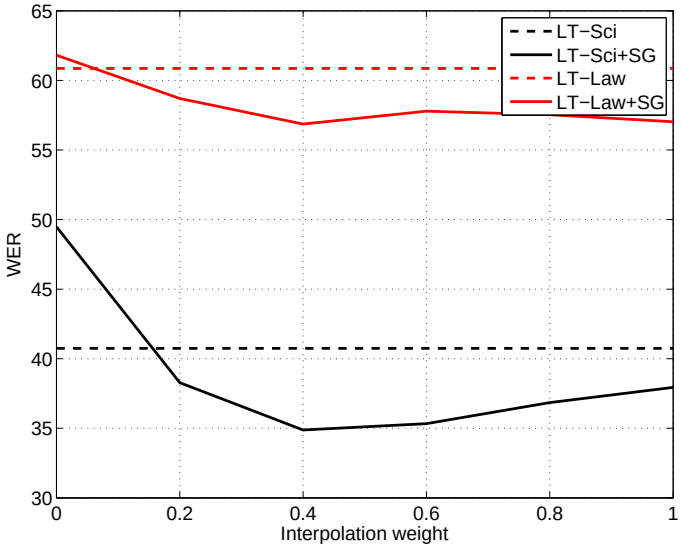

Fig. 2. WER for off-line lecture transcription when trained on *sci* and *law* sources respectively and evaluated on the *sci* LT test set. The dotted lines correspond to language models trained only on the LT training data, and the solid lines represent interpolated results, with the interpolation weight on the horizontal axis.

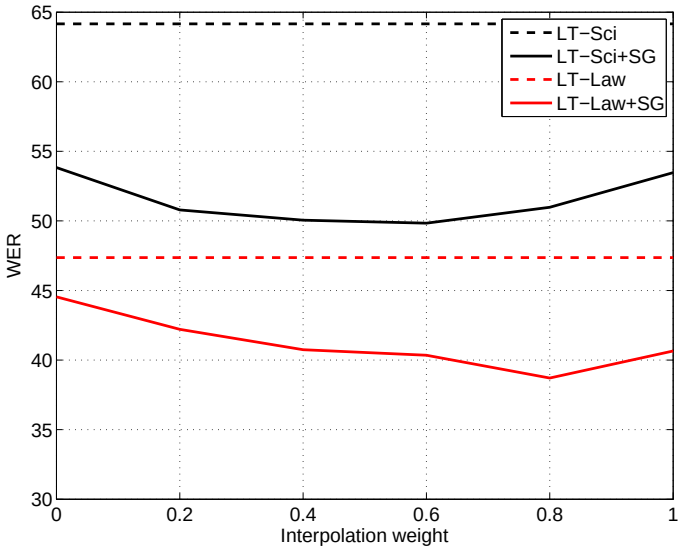

Fig. 3. WER for off-line lecture transcription when trained on *sci* and *law* sources respectively and evaluated on the *law* LT test set. The dotted lines correspond to language models trained only on the LT training data, and the solid lines represent interpolated results, with the interpolation weight on the horizontal axis.

improvement is clear. The most likely sources of such improvement are methods that use the limited acoustic and textual information more efficiently; we therefore believe that the development of such methods should be a priority for further research.

V. Acknowledgment

We would like to thank the NRF (National Research Foundation) for the bursary funding provided throughout the year.

References

[1] 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, 2002, pp. 192–196.

[2] 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*, 2010, pp. 1–6.

[3] 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.

[4] N. J. de Vries, J. Badenhorst, M. H. Davel, E. Barnard, and A. de Waal, "Woefzela - An open-source platform for ASR data collection in the developing world," in *Proceedings Interspeech*, Florence, Italy, August 2011, pp. 3176–3179.

[5] 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, 2007, pp. 2553–2556.

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

[7] C. Munteanu, G. Penn, and R. Baecker, "Web-Based Language Modelling for Automatic Lecture Transcription," in *In Proceedings of the Tenth European Conference on Speech Communication and Technology - EuroSpeech / Eighth INTERSPEECH*, Antwerp, Belgium, 2007, pp. 2353–2356.

[8] I. Trancoso, R. Nunes, H. Moniz, D. Caseiro, and A. I. Mata, "Recognition of Classroom Lectures in European Portuguese," in *INTERSPEECH 2006 - ICSLP (9th International Conference on Spoken Language Processing)*, no. 1. Pittsburgh, PA, USA: ISCA, 2006, pp. 281–284.

[9] A. Jansen and K. Church, "Towards Unsupervised Training of Speaker Independent Acoustic Models," in *Interspeech 2011 (ICSLP 12th Annual Conference of the International Speech Communication Association)*, August 2011, pp. 1693–1692.

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

[11] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book (for HTK Version 3.4)*, march ed. University of Cambridge, 2009, no. July 2000.

[12] 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.

[13] S. F. Chen and J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling," Computer Science Group, Harvard University, Cambridge, Massachusetts, Tech. Rep. TR-10-98, 1998.

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