Active Learning based data selection for limited resource STT and KWS

Thiago Fraga-Silva[1], *Jean-Luc Gauvain*[2], *Lori Lamel*[2],
Antoine Laurent[1], *Viet-Bac Le*[1], *Abdel Messaoudi*[1]

[1]Vocapia Research, 28 rue Jean Rostand, 91400 Orsay, France
[2]CNRS/LIMSI, Spoken Language Processing Group, 91405 Orsay Cedex, France
{thfraga,laurent,levb,abdel}@vocapia.com, {gauvain,lamel}@limsi.fr

Abstract

This paper presents first results in using active learning (AL) for training data selection in the context of the IARPA-Babel program. Given an initial training data set, we aim to automatically select additional data (from an untranscribed pool data set) for manual transcription. Initial and selected data are then used to build acoustic and language models for speech recognition. The goal of the AL task is to outperform a baseline system built using a pre-defined data selection with the same amount of data, the Very Limited Language Pack (VLLP) condition. AL methods based on different selection criteria have been explored. Compared to the VLLP baseline, improvements are obtained in terms of Word Error Rate and Actual Term Weighted Values for the Lithuanian language. A description of methods and an analysis of the results are given. The AL selection also outperforms the VLLP baseline for other IARPA-Babel languages, and will be further tested in the upcoming NIST OpenKWS 2015 evaluation.

Index Terms: active learning, low-resourced STT, KWS.

1. Introduction

This paper describes our recent research in using active learning to select a set of training data in the context of the IARPA-Babel program [8]. The program aims at developing speech-to-text (STT) and keyword spotting (KWS) systems for low-resourced languages. In the context of this work, low-resourced languages are those with a low presence on the Internet, and more generally, limited textual resources especially in electronic form. There is in general little knowledge about the language, with very little or essentially no available audio data and small pronunciation dictionaries (if available). Over the last years there has been growing interest in developing technologies for low-resourced languages as illustrated by the growing popularity of the SLTU workshop series[1] as well as special sessions in major conferences. Some of the approaches range from bootstrapping with models from well-resourced languages to complete self-discovery of linguistic units for unwritten languages (see for example [1, 3, 12, 21, 23, 22]).

The IARPA-Babel program [8] aims to support rapid development of speech technologies for effective keyword search in a variety of languages selected to present challenges at different levels (written scripts & writing conventions, phonological, morphological, dialectal). For each targeted language the program provides a build pack, which contains transcribed speech data, a pronunciation dictionary and a brief descriptive "Language Specific Peculiarities" document [19]. The techniques developed in the program on what are referred to as development languages are also applied to a surprise language as part of the NIST Open Keyword Search Evaluation (OpenKWS13, OpenKWS14, OpenKWS15) [20].

Since the project start, the resources provided within the program for each language have been reduced annually in order to promote the development of techniques which are less dependent on such resources, as their collection and annotation are both time-consuming and costly. A new research direction under exploration is to use an initial STT system trained on a very small amount of transcribed audio data (only one hour) to select additional data to be transcribed. This is defined as the Active Learning (AL) task within the IARPA-Babel program. The basic idea is similar to the techniques used for the unsupervised, semi or lightly supervised acoustic modeling techniques explored over the last decade [11, 14, 15, 25]. However instead of applying the techniques to create approximate transcripts which are directly used for acoustic model training, here the automatic transcripts are used to select a subset of data from a pool of data for which true transcripts will be created.

A variety of criteria were considered for data selection based on knowledge of speech models and experience in training them as described in Section 3. Systems built with the AL based selected data are compared to a baseline system built on a pre-defined data set, called the Very Limited Language Pack (VLLP) condition. Comparisons are performed in terms of word error rate (WER) and actual term weighted value (ATWV). The next section describes the available corpora and methodology. The AL data selection problem is formalized in Section 3. Experimental results and analyses are presented in Section 4 followed by a brief conclusion.

2. Data and methodology

The STT systems were trained on data provided within IARPA-funded Babel program [8]. In this program phase (OP2) systems are being developed for 6 languages: Cebuano, Kazakh, Kurdish, Lithuanian (IARPA-babel304b-v1.0b), Telugu and Tok-Pisin. As a total about 50 hours of transcribed conversational telephone speech are provided for each language. This data is divided into different subsets which are illustrated in Figure 1.

2.1. The Active Learning task

The AL task in the IARPA-Babel program can be described as follows. A pre-defined 1-hour training set is used to build a bootstrap system. This system is used to decode an untran-

[1] http://www.mica.edu.vn/sltu2008 through sltu2014

scribed 29-hour pool data set. Based on the decoding hypotheses and a selection criterion, 2 hours of data are selected from the data pool for manual transcription (transcription recovery). An AL-based STT system is then built using the available 3 hours (initial 1h + selected 2h) of data. The recovered transcriptions cannot be used to perform further iterations of data selection.

In the context of these experiments, the data pool from which the selection is made is already transcribed. So the above procedure simulated by recovering the transcripts from the complete word time-coded corpus. The results reported in this paper were obtained using an internal transcription recovery algorithm. For the OpenKWS15 Evaluation, the transcripts will be provided by NIST. During the development phase, similar results were obtained using our internal recovery algorithm and the transcripts returned by NIST.

2.2. The VLLP baseline condition

Systems built with the AL based data selection were compared to systems built with the VLLP data set. It consists of a pre-defined 3-hour set selected in order to have about the same duration of speech for each speaker represented in a pool of 30 hours of data (see Figure 1). The number of speakers varies from 364 to 399 for the OP2 languages. The VLLP data set includes the 1-hour data set used to bootstrap the AL systems.

2.3. VLLP and AL based system development

For the VLLP and AL tasks, only 3 hours of data are considered to be transcribed. The remainder of the pool data set (27 hours) and additional roughly 40 to 50 hours of untranscribed data for each language were available and could be used for semi-supervised training [11, 25]. For both conditions, the data available from the Year-1 and Year-2 IARPA-Babel program (11 languages) could be used to develop multilingual models.

In addition to the manual transcriptions associated to the 3-hour training data, a textual corpus was available. It consists of texts collected from the Web (Wikipedia, subtitles and other webtexts). This webdata was filtered, normalized and provided to the Babelon team by BBN. The size of the webdata varies between languages from 5.7M to 49M words. For Lithuanian (IARPA-babel304b-v1.0b), about 26M words were available.

Two data sets were used to assess the models. The tuning set, containing 3 hours of speech, was used to optimize the system parameters. The development set, containing about 10 hours of speech, was used to evaluate the systems in terms of speech recognition and keyword spotting (see Figure 1).

2.4. Baseline recognition systems

The baseline STT and KWS systems are described in [9]. For rapid development, all STT systems are based on graphemic pronunciation units and are built via flat start. The acoustic models (AM) are left-to-right 3-state HMMs with Gaussian mixture observation densities, triphone-based and word position-dependent [4]. The models contain about 2k tied-states and 20k mixtures. The models are built using discriminative features produced with a stacked bottle-neck multilayer perceptron and provided to the Babelon team by BUT [7].

Language models (LM) are n-gram based and are trained with the LIMSI STK toolkit. Component models are estimated on the manual transcriptions associated to the acoustic training data and the webtexts. These models are interpolated with coefficients optimized on the TUN data set. Decoding is carried out in a single-pass. A word lattice is generated a 3-gram LM, then a consensus decoding is performed to generate the final hypotheses.

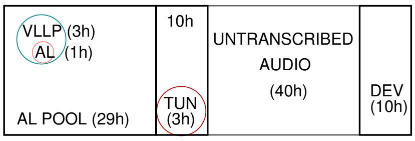

Figure 1: Available data for system training and evaluation in the IARPA-Babel period OP2.

2.5. Keyword search method

The keyword search method used in this work is described in [9]. First, a word and a sub-word consensus network (CN) are generated from decoding lattices [17]. Both CNs are searched to locate all sequences of words and sub-words that correspond to each keyword. Word boundaries are ignored during search.

Keyword hits from both CNs are combined based on time-codes. The keyword scores are then normalized and calibrated using the BBN KST normalization tool [10]. Decision about keeping or ignoring keyword hits is based on a defined threshold. In this work, sub-word units have up to 7 letters (7-grams).

2.6. Performance metrics

Speech recognition system performance is measured using the well-known word error rate (WER) metric. The KWS performance is reported here in terms of the Actual Term-Weighted Value (ATWV) [2, 20]. The keyword specific ATWV for the keyword k at a specific threshold t is computed as:

$$ATWV(k,t) = 1 - P_{FR}(k,t) - \beta P_{FA}(k,t) \qquad (1)$$

where P_{FR} and P_{FA} are respectively the probability of a false reject (miss) and false accept. The constant β mediates the trade off between false accepts and false rejects and is set to 999.9 for the OpenKWS Evaluation.

3. Active Learning

Active Learning based data selection is a research area that has been recently explored for speech and language processing technologies [13, 16, 18, 24]. Kirchhoff et al.[13] identifies at least four applications for which AL could be used: to speed-up system development, for system adaption, for annotation and for system evaluation. The common objective of these applications is to use data selection to meet certain requirements in terms of development time or budget. The selected data set should furthermore contain as much as possible of the information available in the full data set.

Data selection can be formalized as follows. Given a pool data set P, the aim is to select a subset S of P with size $|S| <= |P|$, that maximizes a suitable objective function $f(\cdot)$:

$$S^* = \arg\max\{f(S) : |S| = k, S \in P\} \qquad (2)$$

The scenario defined for the OpenKWS15 is for data annotation. The data set sizes are defined in terms of duration of speech in hours and correspond to $k = 2$ and $|P| = 29$.

In this work, the selection units are speech segments obtained by a Voice Activity Detection (VAD) system. Different monolingual and multilingual VADs were evaluated in order to process the pool data [5]. For the data selection task, the best STT performances were obtained using a VAD based on the time-domain correlation function [6] and trained on multilingual data. It was used in all AL experiments reported here.

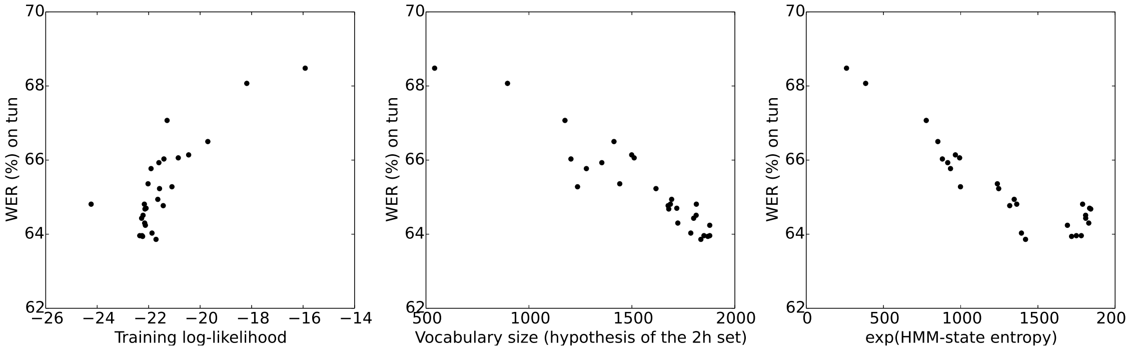

Figure 2: WER vs. training likelihood, vocabulary size of hypotheses and HMM-state entropy for 27 AL-based systems.

Measure	WER	Voc. hyp.	Letter hyp.	LLH train	Conf.	Entropy
WER	1.00	**-0.94**	**-0.84**	**0.82**	0.40	**-0.94**
Voc. hyp.	-	1.00	0.70	**-0.81**	-0.37	**0.95**
Letter hyp.	-	-	1.00	-0.74	-0.12	0.73
LLH train	-	-	-	1.00	0.08	**-0.88**
Conf.	-	-	-	-	1.00	-0.26
Entropy	-	-	-	-	-	1.00

Table 1: Pearson correlation matrix between WER, size of hypothesized vocabulary (Voc. hyp.), number of letters in hypotheses (Letter hyp.), log-likelihood of training data (LLH train), decoding confidence score (Conf.) and HMM-state entropy (Entropy).

3.1. HMM-state entropy criterion

The main selection criterion proposed in this work is the HMM-state entropy, which can be defined as:

$$H = -\sum_{i=1}^{N} \frac{c_i}{C} \cdot log_2 \frac{c_i}{C}, \text{ with } C = \sum_{i=1}^{N} c_i \tag{3}$$

assuming that there exists N acoustic states representing the speech distribution, and where c_i correspond to the number of training vectors associated to the state $i \in [1, N]$. A greedy algorithm was used to solve Equation 3. At each iteration, the utterance giving the highest increase in entropy is selected until the target amount of data is obtained.

Entropy based data selection was already explored in [24]. Authors used entropy based on the distribution of phonemes to uniformly select a subset of a *transcribed* corpus. Here, this approach is extended to the distribution of acoustic model states and for data selection in an *untranscribed* corpus. Intuitively, model states would provide a more general representation of the acoustic space compared to phonemes. In this work, the state labels are obtained from the decoding of the untranscribed corpus using the bootstrap system.

3.2. Other selection criteria

Besides entropy, a variety of other selection criteria were assessed. They have been defined in order to cover different aspects related to acoustics, lexical or pronunciation units and decoding metrics. Selection was performed based on: the signal duration of the segment, the speech density, the utterance data likelihood (normalized by duration or unnormalized), the decoding confidence scores, the number of letters, the number of words and the letter density w.r.t. the signal or the speech duration. For most of these metrics, selection was performed based on the minimum and maximum values (e.g. shortest and longest signal duration).

4. Experimental results

4.1. Correlation analysis

WER and ATWV are the measures of interest for the tasks STT and KWS tasks respectively. Unfortunately, these metrics cannot be used as objective functions for data selection, since the true transcriptions are unknown. A first set of experiments was carried out on Lithuanian aiming to determine which metrics are better correlated with the WER. It was assumed that the WER and ATWV have a strong correlation.

The criteria specified in the previous section were used to build 27 AL-based AMs. The LMs were estimated only on the recovered transcriptions. Various measures were used to analyze the selection and recovered transcriptions. These include the number of words or letters (distinct, total) in the hypotheses and in the recovered references, the training and TUN data likelihood, the HMM state entropy, the decoding confidence scores, the LM perplexity on TUN and the out-of-vocabulary rate.

The curves plotted in Figure 2 show the relation between the TUN WER and likelihood of the selected data, the size of the vocabulary in the decoding hypotheses and the HMM-state entropy for the 27 AL-based systems. Selecting data of low likelihood conducts to better WER results than selecting data of high likelihood. This can be explained by the fact the information provided by the high likelihood data is already present in the initial system, while the manual annotations of data with low likelihood provide more information. The HMM-state entropy and the vocabulary size are correlated with the TUN WER. To reduce the WER, it is necessary to select data with high entropy and a large variety of words. It was observed that the vocabulary size of the hypotheses is strongly correlated with the vocabulary size of the recovered transcripts.

To have an overview of how the various measures are associated, a Pearson correlation matrix was calculated. This matrix

LM used for AL pool decoding	LM used for TUN decoding		
	train(3h)	+web(40k)	+web(100k)
train(3h)	63.9	59.8	59.1
+web(40k)	63.8	59.6	**58.8**
+web(100k)	64.1	59.7	58.9

Table 2: TUN WER(%) with three LMs: one trained on manual transcripts and two with additional webdata (40k and 100k word lists). HMM-state entropy based selection.

Selection criterion	voc. size	trn(3h)	+web(40k)
AL bootstrap (1h)	2.5k	70.0	64.3
VLLP baseline	5.7k	65.6	61.1
Signal duration (max)	5.7k	66.5	61.4
Signal duration (min)	5.4k	65.8	61.2
Confidence score (max)	4.2k	67.1	61.6
Confidence score (min)	6.6k	64.8	60.8
Log-likelihood (max)	5.1k	68.5	64.6
Log-likelihood (min)	6.5k	64.8	60.5
Letter density (max)	**6.9k**	**63.9**	**59.6**
HMM-state entropy	6.4k	**63.9**	**59.6**

Table 3: TUN WER(%) with LM trained manual transcripts and with additional webdata (40k word list). Voc. size: reference transcription vocabulary.

is partially shown in Table 1. Larger numbers in magnitude means stronger correlation. The two measures that correlates at best with the WER are the number of distinct words in the hypotheses (Voc. hyp.) and the Entropy. The likelihood of the data and the number of letters (Letter hyp.) also correlates well with the WER. No significant correlation was observed between the confidence scores and the obtained WER. This is possibly due to the low accuracy obtained with the initial system.

4.2. Impact of webdata on data selection

The experiments described in the previous section were performed using the available 3h data set for acoustic and language modeling. Results suggest that the best criteria tended to maximize the number of distinct words in the selected data. We assessed the impact of adding webdata to the LM used to decode the AL data pool (prior to selection). The underlying premise is that web-based LMs would improve the accuracy of hypothesized transcripts of the pool, thereby helping the data selection. The webdata LM was also used to decode the TUN data.

Web-based LMs were estimated using a 40k and a 100k word list. Table 2 shows the WER obtained with Lithuanian STT systems built using the HMM-state entropy selection. Similar results were obtained with the letter density criterion. Using webdata leads to significant WER reductions on the TUN data. An absolute gain of 4.8% (63.9% vs. 59.1%) is obtained with the 100k-word LM. However, webdata has little impact on data selection based on the AL pool decoding. The largest WER absolute gain obtained was 0.3% (59.1% vs. 58.8%).

4.3. STT and KWS results

Table 3 summarizes the data selection STT experiments performed for Lithuanian. For each selection criterion, the TUN WER obtained with a LM trained only on the transcriptions and with the additional webdata is reported. The vocabulary size of the recovered transcriptions (reference) is also given. The results obtained with the AL bootstrap system and with the VLLP baseline are reported for comparison.

Systems built using the data selection criteria leading to the best STT results on the TUN data were used to decode the DEV data (used here as an evaluation set). KWS was performed on the word/sub-word based consensus networks as described in [9]. Table 4 presents the WER and ATWV results obtained. The ATWV results reported are those obtained after combining word and sub-word keyword hits. 'IV' and 'OOV' refer respectively to the scores obtained on the in-vocabulary and out-of-vocabulary decoding words. Here, a keyword is considered OOV if it contains at least one OOV word.

Condition	WER	ATWV (all/IV/OOV)
VLLP baseline	59.4	0.357 / 0.382 / 0.221
AL, log-likelihood (min)	58.9	**0.387 / 0.407 / 0.268**
AL, letter density (max)	58.8	0.385 / 0.407 / 0.250
AL, HMM-state entropy	**58.4**	0.383 / 0.403 / 0.272

Table 4: WER and ATWV results on the Lithuanian DEV data.

On the DEV data, the reported AL-based systems outperform the VLLP baseline in terms of WER (1% absolute) and ATWV (3% absolute). The best system for STT is not the best for KWS. While the lowest WER is obtained with the entropy criterion (58.4%), the highest ATWV is obtained with the minimum likelihood criterion (0.387). However, we note that ATWV differences among the AL-based systems reported in Table 4 is small (between 0.383 and 0.387).

5. Summary

We explored Active Learning methods using a variety of criteria to select data for manual transcription and STT training. We first analyzed the correlation between various measures to identify the best selection criteria for Lithuanian. It was observed that the HMM-state entropy and the vocabulary size are strongly correlated with the WER. The HMM-state entropy and the letter density criteria, the two recovering the largest vocabularies, led to the best WER, outperforming the baseline VLLP by about 1-1.7% absolute. The best AL-based systems also improved over the baseline ATWV by about 3% absolute.

The data pool from which the selection is taken was already transcribed, so the transcripts are simply recovered from the time-aligned corpus. In real conditions, it might be useful to add constraints on the minimum and maximum segment duration to facilitate the production of manual transcripts.

Similar results have since been obtained for the other IARPA-Babel OP2 languages. Preliminary results indicate that combining selection criteria can improve over the results reported here. These techniques will be applied to the OpenKWS15 Evaluation surprise language.

6. Acknowledgments

We would like to thank our Babelon partners for sharing resources (BUT for the bottle-neck features and BBN for the webdata), and Grégory Gelly for providing the VADs.

This research was in part supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

7. References

[1] L. Besacier, E. Barnard, A. Karpov, T. Schultz, "Automatic speech recognition for under-resourced languages : A survey," *Speech Communication Journal*, vol. 56, pp. 85-100, January 2014.

[2] J. G. Fiscus, J. Ajot, J. S. Garofolo, G. Doddington, "Results of the 2006 spoken term detection evaluation," *ACM SIGIR*, pp. 51–55, 2007.

[3] M. Gales, K. Knill, A. Ragni, S. Rath, "Speech recognition and keyword spotting for low resource languages: BABEL project research at CUED," *SLTU*, 2014.

[4] J. L. Gauvain, L. Lamel and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, no. 1-2, pp. 89–108, 2002.

[5] G. Gelly, J. L. Gauvain. "Minimum Word Error Training of RNN-based Voice Activity Detection," *ISCA Interspeech (submitted)*, 2015.

[6] H. Ghaemmaghami, B. J. Baker, R. J. Vogt, S. Sridharan, "Noise robust voice activity detection using features extracted from the time-domain autocorrelation function," *ISCA Interspeech*, 2010.

[7] F. Grézl, M. Karafiát, "Semi-Supervised bootstrapping approach for neural network feature extractor training," *IEEE ASRU*, pp. 470–475, 2013.

[8] M. Harper, "IARPA Babel Program," `http://www.iarpa.gov/index.php/research-programs/babel`

[9] W. Hartmann, V. B. Le, A. Messaoudi, L. Lamel, J. L. Gauvain, "Comparing decoding strategies for subword-based keyword spotting in low-resourced languages," *ISCA Interspeech*, 2014.

[10] D. Karakos, R. Schwartz, S. Tsakalidis, L. Zhang, S. Ranjan, T. Ng, R. Hsiao, G. Saikumar, I. Bulyko, L. Nguyen, J. Makhoul, F. Grezl, M. Hannemann, M. Karafiat, I. Szoke, K. Vesely, L. Lamel, V.B. Le "Score normalization and system combination for improved keyword spotting," *IEEE ASRU*, pp. 210–215, 2013.

[11] T. Kemp, A. Waibel, "Unsupervised training of a speech recognizer: recent experiments," *ESCA Eurospeech*, pp. 2725–2728, 1999.

[12] T. Kempton, R. Moore. "Discovering the phoneme inventory of an unwritten language: A machine-assisted approach," *Speech Communication Journal*, vol. 56, pp. 152-166, January 2014.

[13] K. Kirchhoff, J. Bilmes, K. Wei, Y. Liu, A. Mandal, C. Bartels. "A submodularity framework for data subset selection," *Technical Report AFRL-RH-WP-TR-2013-0108*, University of Washington, September 2013.

[14] L. Lamel, J.L. Gauvain, G. Adda, "Lightly supervised acoustic model training," *ISCA ITRW Workshop on Automatic Speech Recognition: Challenges for the new Millenium*, pp. 150–154, 2000.

[15] V.B. Le, L. Lamel, A. Messaoudi, W. Hartmann, J.L. Gauvain, C. Woehrling, J. Despres, A. Roy. "Developing STT and KWS systems using limited language resources," *ISCA Interspeech*, 2014.

[16] Y. Liu, K. Wei, K. Kirchhoff, Y. Song, J. Bilmes, "Submodular feature selection for high-dimensional acoustic score spaces," *IEEE ICASSP*, 2013.

[17] L. Mangu, E. Brill, A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Computer, Speech and Language*, 14(4):373-400, 2000.

[18] R. C. Moore, W. Lewis. "Intelligent selection of language model training data," *ACL*, pp. 220–224, 2010.

[19] `http://www.nist.gov/itl/iad/mig/upload/IARPA_Babel_Performer-Specification-08262013.pdf`

[20] NIST Open Keyword Search Evaluation (OpenKWS) `http://www.nist.gov/itl/iad/mig/openkws.cfm`

[21] S. Stücker, M. Müller, Q.B. Nguyen, A. Waibel. "Training time reduction and performance improvements from multilingual techniques on the BABEL ASR task," *IEEE ICASSP*, pp. 6374–6378, 2014.

[22] N. T. Vu, F. Metze and T. Schultz. "Multilingual bottleneck features and its application for under-resourced languages," *SLTU*, Cape Town, South Africa, May 2012.

[23] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," *IEEE ICASSP*, 2014.

[24] Y. Wu, R. Zhang, A. Rudnicky. "Data selection for speech recognition," *IEEE ASRU*, pp. 562–565, 2007.

[25] G. Zavaliagkos and T. Colthurst, "Utilizing Untranscribed Training Data to Improve Performance," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 301-305, 1998.