IMPROVING DATA SELECTION FOR LOW-RESOURCE STT AND KWS

Thiago Fraga-Silva[1], *Antoine Laurent*[1], *Jean-Luc Gauvain*[2],
Lori Lamel[2], *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 extends recent research on training data selection for speech transcription and keyword spotting system development. Selection techniques were explored in the context of the IARPA-Babel Active Learning (AL) task for 6 languages. Different selection criteria were considered with the goal of improving over a system built using a pre-defined 3-hour training data set. Four variants of the entropy-based criterion were explored: words, triphones, phones as well as the use of HMM-states previously introduced in [4]. The influence of the number of HMM-states was assessed as well as whether automatic or manual reference transcripts were used. The combination of selection criteria was investigated, and a novel multi-stage selection method proposed. This method was also assessed using larger data sets than were permitted in the Babel AL task. Results are reported for the 6 languages. The multi-stage selection was also applied to the surprise language (Swahili) in the NIST OpenKWS 2015 evaluation.

Index Terms— data selection, low-resource languages, speech recognition, keyword spotting

1. INTRODUCTION

This paper describes advances in our recent research in using training data selection methods for speech-to-text (STT) and keyword spotting (KWS) system development. This work was performed in the context of the IARPA-Babel program [10] which focuses on low-resource languages. Such languages typically have a low presence on the Internet, with limited textual resources in electronic form and little available knowledge about the language.

Developing speech and language technologies for low-resource languages has been attracting increasing interest in the research community, with dedicated workshops and special sessions at major conferences. A variety of approaches aim to bootstrap models from well-resourced languages to zero-resource languages and to discover linguistic units for unwritten languages [1, 2, 5, 14, 17, 24, 25, 26].

The aim of the IARPA-Babel program is to support the rapid development of speech technologies for effective keyword search in a wide range of languages chosen to cover challenges at different levels (written scripts & writing conventions, phonological, morphological, dialectal). The program sponsors a surprise language evaluation, the NIST Open Keyword Search Evaluation (OpenKWS13, OpenKWS14, OpenKWS15) [22] open to the general community.

A focus of our work is to promote the development of techniques which can easily and quickly be applied to an unknown language. The Active Learning (AL) task within the Babel program explores training a bootstrap STT system on a very small amount of transcribed audio data (only one hour) and using this system to select additional data to be transcribed. The methods investigated are similar to those used for semi or unsupervised acoustic model training [13, 16, 18, 29]. Instead of directly using the approximate transcripts for acoustic model training, here the automatic transcripts are used to select a subset of data from a pool of data for which true manual transcripts will be created.

In [4], we presented some first results addressing the AL task. In that work, a study was conducted in order to provide insight into the correlation between candidate data selection criteria and speech recognition performance. Two criteria appeared to lead to the best overall performance, the HMM-state entropy and the letter density. In this paper, our initial work is extended in various directions.

First, four variants of the entropy-based data selection are explored. Entropy is calculated over the distribution of words or acoustic units, more specifically, phones, triphones and HMM-states. For comparison, selection is made over an *untranscribed* corpus according to the IARPA-Babel AL task premises, and over a *transcribed* corpus like in [28]. It is shown that the quality of the transcriptions (automatic or manual) has little impact on the STT and KWS results for acoustic based selection, but a significant impact on KWS results for word based selection. The influence of the acoustic representation in entropy based selection is assessed. By varying the number of HMM-states that cover the acoustic space, we show that results obtained via HMM-state selection converges

either to those obtained via phone or triphone based selection.

The combination of selection criteria is also investigated in this paper. A novel multi-stage selection method is proposed. The best results are obtained by selecting a sub-pool data set using the letter density criterion, followed by a selection based on the acoustic entropy criterion. Finally, data selection methods are assessed using larger data sets than were permitted in the IARPA-Babel AL task. Both, letter density and acoustic entropy data selection outperform random selection for different data set sizes.

2. DATA SELECTION

Data selection is a research area that has been recently explored for speech and language processing technologies [15, 19, 21, 23, 27, 28]. Kirchhoff et al. [15] identifies at least four applications for which data selection methods are well suited: to speed-up system development, for system adaptation, for data annotation and for system evaluation. Generally speaking, the goal is to select a subset of the data that contain as much as possible of the information available in the full data set. Selection is generally conditioned on a fixed budget, which can be defined in terms of human effort, development time, processing time, data set size or any other requirement.

Formally, data selection can be considered as a constrained optimization problem. 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 $f(\cdot)$:

$$S^* = \arg\max\{f(S) : |S| = k, S \in P\} \quad (1)$$

Within the IARPA-Babel AL task, selection is used to simulate manual annotation. Thus, the objective function has to be calculated over any feature or combination of features that can be obtained from acoustic analysis or as a result of speech decoding (e.g. amount of speech, confidence measures, data likelihood, entropy, number of hypothesized words). Furthermore, data set sizes are defined in terms of duration of speech in hours and correspond to $k = 2$ and $|P| = 29$.

In all experiments reported here, the selection units are speech segments obtained by a voice activity detection (VAD) system [7] based on the time-domain correlation function [8] and trained on multilingual data.

3. THE IARPA-BABEL ACTIVE LEARNING TASK

The STT and KWS systems were trained on data provided within the IARPA-funded Babel program [10]. In this phase of the program (OP2), systems were built for 6 *development* languages (Cebuano, Kazakh, Kurdish, Lithuanian, Telugu and Tok-Pisin) and the *surprise* language (Swahili)[1]. The goal of the AL task is not to have the most accurate system, but to obtain the best improvements over a baseline trained in similar conditions (amount of data, acoustic features, vocabulary size, language models, decoder, etc).

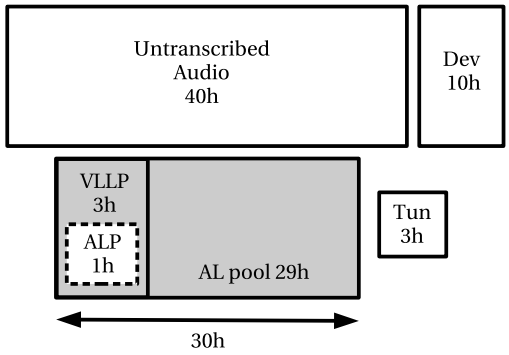

Fig. 1: Available data for system training and evaluation in the IARPA-Babel period OP2.

3.1. Corpus

A total of about 50 hours of transcribed conversational telephone speech were provided for each language. This corpus is divided into different subsets which are illustrated in Figure 1. For the AL task, 30 hours of speech are considered as the full training data set within which selection is made. One hour of data is fixed. A 3-hour TUN data set is used to optimize system parameters, while a 10-hour DEV data set is used only to assess the models (as a test set). Additional 40 hours of untranscribed data were available for each language and could be used for semi-supervised training [13, 29]. Data available from the Year-1 and Year-2 IARPA-Babel program (11 languages) could be used to develop multilingual models.

The baseline system is built upon a pre-defined 3-hour set, known as the Very-Limited Language Pack (VLLP). This data set was 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 VLLP baseline includes the fixed 1-hour data set common to the AL based systems.

In addition to the manual transcriptions of the 3-hour training data, a textual corpus was available. It consists of texts collected from the Web (Wikipedia, subtitles and other webtexts). The version of the webtexts used here was filtered, normalized and provided to the Babelon team by BBN [30].

3.2. Data selection protocol

The IARPA-Babel AL task is depicted in Figure 2. A predefined 1-hour training set is used to build a bootstrap system. This system is used to decode an untranscribed 29-hour pool

[1]Language Packs: Cebuano (IARPA-babel301b-v2.0b), Kazakh (IARPA-babel302b-v1.0a), Kurmanji (IARPA-babel205b-v1.0a), Lithuanian (IARPA-babel304b-v1.0b), Telugu (IARPA-babel303b-v1.0a), Tok-Pisin (IARPA-babel207b-v1.0b) and Swahili (IARPA-babel202b-v1.0d)

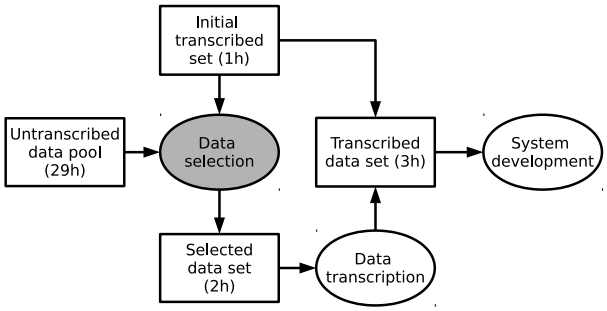

Fig. 2: Description of the IARPA-Babel Active Learning task.

data set, previously processed by a VAD system. Based on selection criteria derived from the audio analysis and decoding hypothesis, 2 hours of data are selected from the data pool for manual transcription. An AL-based STT system is then built using the available 3 hours (initial 1 hour + selected 2 hours) of data. The available webtexts and untranscribed acoustic data can also be used in system development.

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

3.3. Performance metrics

Speech recognition system performance is measured in terms of word error rate (WER), calculated as the distance edition between the recognition hypothesis and the reference. KWS performance is reported in terms of the Actual Term-Weighted Value (ATWV) [3]. The keyword specific ATWV for the keyword k at a specific threshold t is computed as:

$$ATWV(k,t) = 1 - P_{FR}(k,t) - \beta P_{FA}(k,t) \qquad (2)$$

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.

4. SYSTEM DESCRIPTION

The baseline STT and KWS systems were built using the same methods as described in [4] and [11]. All STT systems are based on graphemic pronunciation units and are built via flat start. Acoustic models are triphone-based left-to-right 3-state HMMs with Gaussian mixture observation densities [6]. Both, word position dependent and word position independent models are generated. About 2k tied-states and 20k mixtures are used. Acoustic features are discriminatively trained. They are extracted using a multilingual stacked bottle-neck multilayer Perceptron and were provided to the Babelon team by BUT [9].

Back-off n-gram based language models (LM) are used. They were estimated using the LIMSI STK toolkit. First, component models are estimated on the manual transcriptions and the webtexts. They are then interpolated with coefficients optimized on the TUN data set. The recognition vocabulary is automatically selected based on unigram probabilities. Vocabulary sizes range between 40k and 60k across the 6 development languages. Decoding is carried out in a single-pass. First, a 2-gram LM is used to generate a word lattice, which is rescored with a 3-gram LM. Then a consensus decoding is performed to generate the final hypothesis.

The keyword search method used here is described in [11]. First, a word and a sub-word decoding are performed. A consensus network (CN) is generated from the decoding lattices [20] for each case. 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 the hypothesized time-codes. The keyword scores are then normalized and calibrated using the BBN KST normalization tool [12]. Decision about keeping or ignoring the keyword hits is based on a defined threshold (0.5 in our experiments). Sub-word units are automatically discovered based on an iterative procedure for optimizing the text perplexity [11].

5. ENTROPY-BASED DATA SELECTION

An HMM-state entropy-based selection was used in [4] for the IARPA-Babel AL task. This criterion was also found to be one of the best in terms of WER for the 6 development languages. In this paper, this criterion is extended to the distribution of other speech units, namely words, phones and triphones. Generally, the entropy function can be defined as:

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

where c_i corresponds to the number of training instances associated to the speech unit $i \in [1, N]$, N being the number of units representing the speech distribution. A greedy algorithm is applied to maximize the entropy function shown in Equation 3. At each iteration, the utterance giving the highest increase in entropy is selected until the target amount of data is obtained.

Phone and word entropy based selection over a transcribed corpus was previously explored in [28]. In [4] the HMM-

Unit	Manual		Automatic	
	WER	ATWV	WER	ATWV
Word	59.3	**0.377**	59.0	0.360
Phone	59.4	0.349	59.0	0.348
Triphone	58.0	0.373	**58.0**	**0.370**
States	**57.7**	0.369	58.1	0.354

Table 1: Lithuanian dev WER(%) and ATWV. Comparison of entropy based data selection guided by manual or automatic transcriptions. The VLLP baseline is 58.7% (0.351).

state entropy was already used over an automatically transcribed corpus.

5.1. Selecting over manual or automatic transcriptions

In this section, entropy selection based on the four proposed speech units is addressed. Selection is performed on the available untranscribed corpus. In this case, the speech unit labels are provided by the bootstrap system. To assess the impact of the quality of the speech recognizer, data selection is also performed using the corresponding manually transcribed corpus. Here, manual transcriptions are considered to be the output of an error-free recognition system. In this case, a forced alignment is performed to obtain the labels for the acoustic based units (phones, triphones, HMM-states).

This comparison was performed on the Lithuanian language pack. In these experiments, only full-word based keyword search was performed. Results are shown in Table 1.

Entropy selection is little affected by the quality of the transcriptions. Somewhat surprisingly, better STT performances are obtained when selection is guided by automatic transcriptions. However, it is important to note that all systems are trained in a supervised manner (only the data selection changes). In terms of ATWV, selection based on phone and triphone units seem to be more robust w.r.t. the quality of transcriptions. The biggest difference in ATWV is observed for word based selection: 0.377 over manual and 0.360 over automatic transcriptions. Selection is somehow misguided by the large number of wrongly recognized words present in the automatic transcriptions. For information, the WER with the bootstrap system is about 65% on the dev data.

5.2. Acoustic space coverage

In [4], we argued that HMM-states should be a better representation of the acoustic space in comparison to phones, and therefore should lead to better data selection results. To a certain extent, results from Table 1 support that claim: HMM-states (and triphone) based selection outperforms phone based selection.

The impact of the acoustic space coverage on entropy selection was assessed for Lithuanian and Tok-Pisin languages. Here, coverage is interpreted as the number of classes used in selection, which can be phones (dozens of units), triphones (about 4000 in these experiments) or HMM-states. The number of tied-states was varied from about 100 to 10k by changing the state clustering threshold. Table 2 summarizes the WER and ATWV results obtained.

Language		100	1k	2k	5k	7k	10k
Lit	WER	58.4	58.3	**58.1**	**58.1**	**58.1**	58.2
	ATWV	0.348	0.358	0.354	0.362	**0.364**	0.363
Tok	WER	49.7	49.5	49.6	49.5	**49.4**	49.6
	ATWV	0.296	0.298	0.302	0.304	0.303	**0.305**

Table 2: Dev WER(%) and ATWV for acoustic entropy selection as a function of the number of HMM-states. 'Lit': Lithuanian; 'Tok': Tok-Pisin.

For both languages, STT performance changes by only 0.3% absolute between the worst (100 states) and best (2k to 7k states) cases. Reducing the acoustic space to 100 states also degrades the KWS performance, with ATWV results approaching those of phone based selection, which are 0.348 for Lithuanian and 0.295 for Tok-Pisin. The best ATWV performances are obtained with a larger number of states, 7k for Lithuanian and 10k for Tok-Pisin. These results are however worse than those obtained with triphone based selection, 0.370 and 0.310 respectively for Lithuanian and Tok-Pisin. This slight difference might be due to the fact that triphone units may encode co-articulation information in addition to acoustic characteristics.

5.3. Extension to all the development languages

Data selection was performed for the 6 development languages using the four variants of the entropy selection. In these experiments, the ATWV is calculated on the combined hits of word and sub-word keyword search. For all languages, 2000 states were used for HMM-state entropy selection. Results are summarized in Table 3.

Except for Telugu, significant improvements over the VLLP baseline were obtained with the entropy based selection methods. Driving selection by triphone or HMM-state units led to good STT and KWS performances across all languages. In particular, selection based on triphones obtains the best or close to best results for each language. We note though that WER and ATWV differences are small among the entropy based systems.

6. COMBINING SELECTION METHODS

In our previous work [4], the acoustic entropy, the letter density and the data likelihood were found to be the best data selection criteria among those explored. Here, a method for combining some of the criteria is proposed. A multi-stage selection was performed as follows. First, a certain criterion *A*

Language	VLLP	Word	Phone	Triphone	States
Cebuano	65.9	65.5	65.8	64.9	**64.6**
Kazakh	66.2	65.3	64.4	**64.0**	64.2
Kurmanji	72.0	71.9	71.2	**70.4**	70.7
Lithuanian	58.7	59.0	59.0	**58.0**	58.1
Telugu	**77.3**	78.1	77.9	78.1	78.5
Tok-Pisin	51.3	50.2	49.6	**49.5**	49.6

(a) Dev WER.

Language	VLLP	Word	Phone	Triphone	States
Cebuano	0.243	0.260	0.255	**0.265**	0.261
Kazakh	0.291	0.288	0.282	**0.294**	**0.294**
Kurmanji	0.169	**0.187**	0.184	0.183	0.182
Lithuanian	0.373	0.399	0.391	**0.403**	0.398
Telugu	0.168	**0.174**	0.173	0.170	0.169
Tok-Pisin	0.296	0.298	0.295	**0.310**	0.302

(b) Dev ATWV.

Table 3: Dev WER(%) and ATWV with systems built using entropy based data selection with different units and the VLLP baseline.

Language		VLLP	AL (size of sub-pool in hours)			
			5	7	10	15
Lit	WER	58.7	57.9	**57.8**	**57.8**	58.1
	ATWV	0.351	0.363	0.362	**0.364**	**0.364**
Tok	WER	51.3	**48.9**	49.2	49.2	49.6
	ATWV	0.296	0.304	**0.315**	0.301	0.303

Table 4: Multi-stage data selection. A sub-pool is selected using letter density. Final selection is made via triphone based entropy selection. 'Lit': Lithuanian; 'Tok': Tok-Pisin.

is used to select a sub-pool of data, from which a 2-hour data set is selected using another criterion *B*.

The correlation study developed in [4] was used to get insight into which criteria would be best candidates for combination. The measures with the largest correlation to the WER were the vocabulary size and acoustic entropy (> 0.9) and the letter density and data likelihood (> 0.8). Among the possible combinations of these four measures, the pair acoustic entropy and letter density has the weakest correlation (0.73). Intuitively, these two latter should combine well.

In preliminary tests, combinations of likelihood, letter density and acoustic entropy were tested for multi-stage selection on Lithuanian data. The best STT performance was obtained by doing a pre-selection via letter density and applying acoustic entropy for the final selection. The other combinations were not investigated further.

The first experiments with the proposed multi-stage selection method were performed to assess the impact of the sub-pool size. Sub-pool data sets having 5 to 20 hours were selected via letter density. Within each sub-pool, a 2-hour data set was selected via triphone based entropy selection. Generally, the best performances were achieved for sub-pool data set sizes of about 7 to 10 hours of speech. However, the optimal case depends on the language. An extract of the WER and ATWV results for Lithuanian and Tok-Pisin is shown in Table 4. For Lithuanian, the optimal sub-pool size is around 10h, while for Tok-Pisin, it is around 7h.

Language	VLLP	Triphone	Density	Multi-stage
Cebuano	65.9	64.9	**64.4**	64.5
Kazakh	66.2	**64.0**	65.1	64.4
Kurmanji	72.0	70.4	70.4	**70.3**
Lithuanian	58.7	58.0	58.3	**57.8**
Telugu	77.3	78.1	77.7	**77.2**
Tok-Pisin	51.3	49.5	49.5	**48.9**

(a) Dev WER.

Language	VLLP	Triphone	Density	Multi-stage
Cebuano	0.243	0.265	0.260	**0.268**
Kazakh	0.291	0.293	0.298	**0.299**
Kurmanji	0.169	0.183	0.182	**0.193**
Lithuanian	0.373	0.403	0.400	**0.413**
Telugu	0.168	0.170	0.175	**0.180**
Tok-Pisin	0.296	0.310	0.298	**0.315**

(b) Dev ATWV.

Table 5: Dev WER(%) and ATWV with the VLLP baseline and systems built using triphone, letter density and multi-stage data selection.

The multi-stage selection method was assessed for all development languages. The proposed method was compared to the VLLP baseline and selection driven by the triphone entropy and the letter density criterion. Results are summarized in Table 5. These results are directly comparable to those shown in Table 3. For multi-stage selection, the size of the sub-pool was fixed to 7h for all languages.

All the selection methods shown in Table 5 outperform the VLLP baseline in terms of ATWV. In terms of WER, this is the case for all languages, except Telugu. The proposed multi-stage selection method gives the best overall performances in terms of ATWV for all languages. It also obtains the best WER results for Kurmanji, Lithuanian, Telugu and Tok-Pisin.

7. SELECTING MORE DATA

Another set of experiments was performed to assess the behavior of the different methods when selecting more than 3 hours from the data pool. Figure 3 shows the WER and ATWV obtained using HMM-state entropy (with 2k states), triphone entropy and letter density for Lithuanian. They are compared to a baseline random selection. For this baseline, the speech segments of the pool set were randomly sorted and

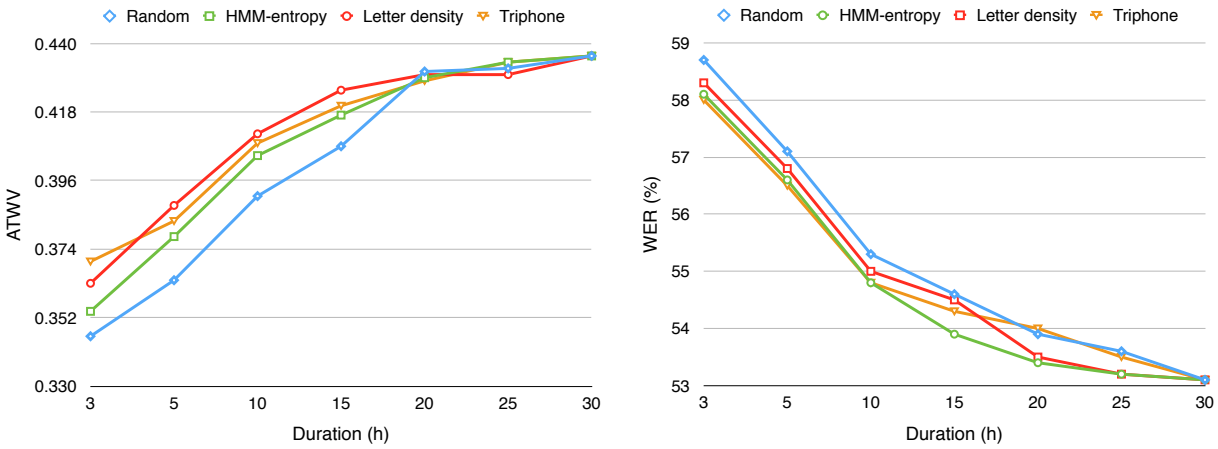

Fig. 3: ATWV and WER in Lithuanian when selecting more data

the first n-segments used for training.

As can be expected in all cases, training the systems on more data reduces the WER and increases the ATWV. In terms of ATWV, the triphone selection gives the best result with 3 hours, but letter density works better with larger amounts (5 to 20 hours). It is interesting to note that 3 hours of entropy selection gives better ATWV results than 5 hours chosen randomly. The ATWV differences are very small across the different criteria with more than 20h of data (from 66% of the full data set).

8. SUMMARY

This paper extended our research work in data selection for low resource languages [4]. Various selection techniques were explored in the context of the IARPA-Babel Active Learning task for 6 languages with the goal of outperforming a baseline STT and KWS system built on 3 hours of a pre-defined data set.

Entropy based selection was extended to word, phone and triphone units and compared to HMM-state based selection introduced in our previous work. Triphone and HMM-state selection were shown to significantly outperform the baseline system and, in most cases, gave better results than word or phone units. Furthermore, entropy selection based on acoustic units were shown to be robust with respect to the quality of the transcription hypothesis. The influence of the acoustic space coverage on data selection was also assessed. This was done by varying the number of tied-states for entropy based selection. Better results were obtained for larger acoustic spaces ($>$7k states and triphones) in comparison to reduced spaces ($<$1k states and phones).

A novel method for combining selection criteria was proposed. The best ATWV performances for all languages were obtained by performing a multi-stage selection: the letter density criterion is used to select a sub-pool of data, from which final selection is made via the triphone entropy criterion. Furthermore, three of the explored criteria (letter density, HMM-state entropy and triphone entropy) were used to select larger data sets than were permitted in the Babel AL task. They all outperformed a random selection baseline in terms of WER and ATWV, especially for smaller training sets.

9. 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.

10. REFERENCES

[1] X. Anguera, L.J. Rodriguez-Fuentes, I. Szaoke, A. Buzo, F. Metze, M Penagarikano, "Query-by-Example Spoken Term Detection Evaluation on Low-Resource Languages", *SLTU 2014*, 2014.

[2] 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.

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

[4] T. Fraga-Silva, J-L. Gauvain, L. Lamel, A. Laurent, V-B. Le, A. Messaoudi, "Active Learning based data selection for limited resource STT and KWS," *ISCA Interspeech*, 2015.

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

[6] 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.

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

[8] 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.

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

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

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] A. Laurent, W. Hartmann, L. Lamel, "Unsupervised Acoustic Model Training for the Korean Language," *ISCA Interspeech*, 2014.

[18] 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.

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

[20] 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.

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

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

[23] C. Ni, L. Wang, H. Liu, C.C. Leung, L. Lu, B. Ma; "Submodular data selection with acoustic and phonetic features for automatic speech recognition," *IEEE ICASSP*,

[24] 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.

[25] 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.

[26] 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.

[27] K. Wei, Y. Liu, K. Kirchhoff, J. Bilmes, "Unsupervised submodular subset selection for speech data," *IEEE ICASSP*, 2014.

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

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

[30] L. Zhang, D. Karakos, W. Hartmann, R. Hsiao, R. Schwartz and S. Tsakalidi, "Enhancing Low Resource Keyword Spotting with Automatically Retrieved Web Documents," *ISCA Interspeech*, 2015.