ACTIVE LEARNING FOR LOW-RESOURCE SPEECH RECOGNITION: IMPACT OF SELECTION SIZE AND LANGUAGE MODELING DATA

Ali Raza Syed[⋆] *Andrew Rosenberg*[†] *Michael Mandel*[⋆]

[⋆] The Graduate Center, CUNY, New York, NY, USA
[†]IBM TJ Watson Research Center, Yorktown Heights, NY, USA

ABSTRACT

Active learning aims to reduce the time and cost of developing speech recognition systems by selecting for transcription highly informative subsets from large pools of audio data. Previous evaluations at OpenKWS and IARPA BABEL have investigated data selection for low-resource languages in very constrained scenarios with 2-hour data selections given a 1-hour seed set. We expand on this to investigate what happens with larger selections and fewer constraints on language modeling data. Our results, on four languages from the final BABEL OP3 period, show that active learning is helpful at larger selections with consistent gains up to 14 hours. We also find that the impact of additional language model data is orthogonal to the impact of the active learning selection criteria.

Index Terms— Active Learning, Data Selection, Automatic Speech Recognition, Low Resource Languages, Language Model

1. INTRODUCTION

Automatic speech recognition (ASR) requires transcribed speech for training. High quality speech transcription is expensive and time-consuming, especially for low resource languages where expert transcription services are limited. Active learning is a general technique for training a classifier that seeks to identify those unlabeled data points that would have the most benefit for performance if labels were available. For ASR, this involves identifying candidate speech segments for transcription.

OpenKWS and IARPA BABEL are public and government sponsored programs, respectively, that investigate ASR and keyword search (KWS) on low resource languages. In the 2015 evaluations, the amount of training material was dramatically limited to 3 hours. In addition to a fixed, sponsor-defined training set, participants were given an opportunity to develop active learning approaches for selecting a 2 hour subset of data, and adding this to a fixed, predefined 1-hour "seed" set. This evaluation and shared task showcased a number of active learning approaches [1, 2, 3] highlighting the efficacy of submodular functions (Section 3) [4, 5, 6].

In this work, we extend the investigation of active learning on speech recognition performance beyond the constraints of the OpenKWS and BABEL evaluations. Specifically we investigate two qualities. First, active learning is only useful when there is real limit to the amount of data that can be selected for labeling. If all of the available, unlabeled data can be labeled, then the active learning selection criteria makes no difference. We evaluate selections of 3 hours, 5 hours, 10 hours and 15 hours to identify the point of diminishing returns for this task. All experiments are performed on IARPA BABEL data; specific information about the distributions can be found in Table 1.

Second, in the OpenKWS and BABEL evaluations, the selected, transcribed material was used for both acoustic and language modeling. However, it is much easier to collect text data to augment a language model than it is to collect and transcribe additional speech for acoustic modeling. One hypothesis explaining the limited performance of ASR trained on 3 hours of data, and the success of active learning over random selections, is that it is due to an impoverished language model. To test this, we augment each language model with material collected from the web and evaluate the impact on recognition performance (Section 4.4).

Language	Build Pack	Pool	WER (FLP)
Amharic	IARPA-babel307b-v1.0b	16.72 h	56.60%
Igbo	IARPA-babel306b-v2.0c	18.97 h	75.60%
Mongolian	IARPA-babel401b-v2.0b	17.60 h	73.40%
Pashto	IARPA-babel104b-v0.4bY	20.11 h	60.20%

Table 1. *Languages, build packs, total hours of segmented audio in selection pool, and word error rate for Full Language Pack (FLP).*

2. RELATED WORK

A number of approaches have been proposed for active learning in speech recognition. Supervised techniques involve the use of transcripts for the process of selecting data. The typical process uses the seed transcripts to train an ASR and decode the audio selection pool. ASR-derived measures, often confidence scores, are used to estimate the likelihood of an utterance's improving the ASR model with high resource languages (e.g. [7, 8, 9, 10]). Our experiments focus on unsupervised techniques which forego the use of any base transcripts and use acoustic features or side information to estimate the likelihood an utterance's importance in improving the ASR model. One advantage of these techniques is faster training of a base ASR model by avoiding the time- and labor-intensive steps of base-level transcription or ASR training and decoding for selection purposes.

There has been growing interest in submodular optimization for data selection and active learning in speech. The facility location function, with a Fisher Kernel for pairwise similarity, was used by [11] for phone recognition in TIMIT, and in low resource ASR training by [5] augmented with a diversity reward function. Feature-based functions have been used by [12] for ASR training with Switchboard and Fisher, while multi-layer feature-based functions were introduced by [13] for phone recognition in TIMIT. Feature-based function have also been applied in low resource settings to select acoustically diverse subsets [6], minimize divergence to a target set [14], and augmented with length-normalization to remove bias for long utterances [4]. Our approach focuses on the application of multi-layer feature-based functions to low resource speech recognition and investigates the effect of

feature encodings on the selections and performance.

3. SUBMODULAR FUNCTIONS AND SUBSET SELECTION

A submodular function [15] f is a real-valued function over sets satisfying the following property for any $S \subseteq S', u \notin S'$:

$$f(S \cup \{u\}) - f(S) \geq f(S' \cup \{u\}) - f(S'). \quad (1)$$

Submodularity captures the property of diminishing returns: the marginal benefit of adding an element u to a set S' does not exceed the marginal benefit of adding u to any subset of S'. f is also monotone non-decreasing if, for any $S \subseteq S'$, $f(S) \leq f(S')$. In the subset selection problem with a cardinality constraint, $|S| < K$, we wish to maximize the value of our selected subset, $f(S)$, subject to S having at most K elements. Although this problem is NP-hard, if f is non-decreasing and submodular, a greedy algorithm ensures a near-optimal subset [16]. The greedy scheme always selects the element with the highest marginal benefit: $f(S \cup \{u\}) - f(S)$. Active learning is often concerned with subset selection under a budget constraint. Here, each element $v \in S$ is associated with a cost, $c(v)$, and the cost of the subset may not exceed a budget B, i.e. $\sum_{v \in S} c(v) \leq B$. In this case, a near-optimal result is guaranteed by the cost-benefit greedy algorithm [17], which always selects the element with the highest marginal benefit cost ratio: $\frac{f(S \cup \{u\}) - f(S)}{c(u)}$.

In this section, we discuss two classes of functions, previously applied in unsupervised data selection for low-resource languages.

3.1. Facility location with diversity reward

The facility location function (2) measures the similarity of a selection S to the remainder of the whole set U [11]:

$$f(S) = \sum_{u \in U} \max_{s \in S} w_{u,s}. \quad (2)$$

Maximizing f above tends to yield elements which are highly representative of the larger pool. One issue is that this function tends to oversample central data points possessing the most common features. The facility location with diversity reward [5] mitigates this by regularizing the facility location objective with a reward based on cluster membership [18]:

$$f(S) = \sum_{u \in U} \max_{s \in S} w_{u,s} + \lambda \sum_{i} \delta(S \cap C_i \neq \emptyset). \quad (3)$$

$C_i, i = 1, \ldots, k$ are clusters which non-disjointly partition U. The second term in Equation 3 counts the number of unique clusters for which the elements of S may claim membership, thus preferring elements possessing a wider range of features. Maximizing f yields a selection which is representative, while selecting from diverse regions of the feature space.

3.2. Two-layered feature based functions

Although the facility location with diversity reward performs well among a variety of unsupervised methods of selection [5], it requires the computationally expensive task of constructing a pairwise similarity graph over the entire dataset. Feature-based submodular functions [19], in contrast, take the form:

$$f(S) = \sum_{h \in H} w_h g_h(m_h(S)). \quad (4)$$

Here, H is a set of features and m_h assigns a relevance score for feature $h \in H$ to set S by summing up the relevance of feature h for each element of S: $m_h(S) = \sum_{s \in S} m_h(s)$. Typically, m_h measures mass or degree of feature h present in sample s. w_h is a feature-specific weight, and g_h may be any non-negative, non-decreasing concave function, such as the logarithm or square root.

Feature-based functions allow a variety of features to be considered for subset selection and tend to promote representative and diverse selections. Although feature-based functions account for interactions between elements of a set, they do not consider interactions between *features*. This has the disadvantage of selecting items even if they posses redundant information. To overcome this, [13] proposed two-layered feature-based submodular functions of the form:

$$f(S) = \sum_{h \in H} g_H \left(\sum_{l \in L} w_{hl}\, g_L\left(m_l(S)\right) \right). \quad (5)$$

Here, L is a low-level feature space, while H is a high-level feature space (e.g. a set of meta-features), such that $\dim(L) > \dim(H)$, where $\dim(\cdot)$ measures the dimensionality of a feature space. Then, w_{hl} measures the interaction between features $h \in H$ and $l \in L$, while m_l measures the relevance score for a set for feature $l \in L$. g_H and g_L are non-negative, non-decreasing concave functions specific to each feature space, as above.

4. METHODS

All ASR experiments were performed using IBM Attila [20] We train both GMM and DNN acoustic models. The GMMs are speaker-independent with 1000 context-dependent (CD) states and 6000 mixture components. The DNNs concatenate 9 input frames, contain 5 hidden layers of 1024 units, a bottleneck layer with 128 units, and 1000 outputs corresponding to the same 1000 CD states. The GMM architecture was optimized for 3 hour selection sets, while the DNN architecture was optimized for systems trained with 40 hours of training data. As such both of these may be sub-optimal at some selection sizes. We chose to keep these architectures fixed to focus on the impact of active learning.

We follow the IARPA BABEL active learning scenario: we are given one hour of transcribed seed data, and a pool of untranscribed audio data for selections. The selection pool was segmented using a Voice Activity Detector (VAD) with frame energy based tresholding from the Spear toolkit [21]. We found the automatic segmentation to be less reliable for segments with duration under one second. Thus, we only considered segments greater than one second in duration.

4.1. Random selection

For a baseline, we use a heuristic based on the one used by IARPA BABEL and NIST (OpenKWS) for selecting 3-hour very limited language packs (VLLP). For each conversation, start at the midpoint and select the closest segment. In successive iterations, select segments by alternately moving toward the beginning and end of the conversation and skipping over segments shorter than one second, or longer than twenty seconds. Selections are made round-robin over all conversations until the required total duration has been met.

4.2. Speech Rate selection

We also performed selections based on speech rate. Selecting utterances with higher speaking rates should yield higher concentration of phones or words for acoustic model and language model training.

Language	Token Count	Type Count
Amharic	10.8M	908.6K
Igbo	1.6M	48.5K
Mongolian	115.9M	1.7M
Pashto	100.5M	1.8M

Table 2. *Token and type counts of web data added to LM.*

We used a syllable nuclei detection algorithm [22] in AuToBI [23] to estimate speech rate.

4.3. Submodular selection

For acoustic features, we used 62-dimensional multilingual features extracted from a deep neural network [24] trained on 28 languages in the BABEL corpus. For our data selection procedures, we discretized the acoustic feature space by learning a k-means codebook $C_1, C_2, \ldots, C_k$ over the utterances with an encoding function q. In any utterance u with m frames, each frame is assigned to its centroid from the codebook. We encode the utterance as the vector $q(u) = [c_i(u) \cdots c_k(u)]$, where $c_i(u)$ counts the occurrence of centroid C_i in u. To instantiate the two-layered feature-based function, we consider two approaches.

Feature encoding with TF-IDF normalization. For the low level feature space L, we learn a k-means clustering with $k = 1024$; for the high level feature space H, we use $k = 128$. The utterances are encoded using both codebooks. To measure interactions between H and L, w_{hl} counts co-occurrence of features $h \in H$ and $l \in L$ across all utterances. The relevance of feature $m_l(u)$ is measured as the TF-IDF normalized count of feature l in u.

Feature encoding based on word2vec skip-gram model. We learn a clustering with $k = 1024$ to encode the utterances as described above. We then employ the word2vec algorithm with a continuous skip-gram model [25] to determine a different distributed representation for the encoding. The skip-gram model uses each frame to predict its surrounding context frames and we expect the resulting encoding to take frame level contexts into account. We learn two different models H and L to produce two sets of encodings, with $d(H) = 42$ and $d(L) = 1024$. The interactions between H and L are measures by computing the covariance features $h \in H$ and $l \in L$ across all utterances, i.e. $w_{hl} = \text{cov}(h, l)$.

4.4. Addition of web data

Having limited amounts of transcribed training data for an ASR limits the performance of both the acoustic and language models. In our experiments, active learning based on acoustic features helps acquire transcribed data with the best potential impact on the acoustic model. However, the resulting language model is tightly coupled to the acoustic data which limits its performance on future unseen data, especially out of vocabulary data. There is a vast amount of textual data on the web and although harvesting such data presents its own set of challenges, the acquisition of additional web data has relatively low marginal costs. We used web data, which had been collected based on the system described in [26], to augment our language model data during training and investigate its impact on system performance. Table 2 shows relevant statistics concerning the web data used to augment the language models.

Language	Hrs	Baseline	Base+Web	SM	SM+Web
Amharic	2	67.2	66.5	65.3	64.6
	4	64.3	63.9	62.8	62.0
	9	60.8	60.3	60.1	59.7
	14	59.5	59.2	58.6	58.1
Igbo	2	77.3	77.1	77.0	76.8
	4	77.1	76.8	76.7	76.5
	9	76.1	75.8	75.4	75.4
	14	75.4	75.4	75.3	75.3
Mongolian	2	76.0	75.4	75.9	75.3
	4	76.4	75.6	74.9	74.2
	9	74.7	74.0	74.1	73.6
	14	74.4	73.8	73.2	72.9
Pashto	2	66.1	65.9	66.0	65.7
	4	64.8	64.6	64.6	64.3
	9	62.6	62.2	62.0	61.7
	14	61.7	61.5	61.0	60.9

Table 3. *Performance with augmented language model.*

4.5. Adapting to the target audio

Evaluation audio may not follow the same feature distribution as the training audio. To measure the impact of this on our selection process, we modified the two layered TF-IDF feature based function by learning clusterings on acoustic features derived from the *evaluation* audio. These encoding functions were then used to encode the training audio data as above. This is an unsupervised adaptation of the training data selection method that is aware of the evaluation audio, but requires no transcription of the evaluation speech.

5. RESULTS

To examine the difference between feature encodings for the SM method, we conducted experiments on Mongolian language data. The results (Fig. 3) show that selection with the two-layer function using TF-IDF feature encodings perform best at all selection points. Addition of web data to this method also improves the results. However, word2vec skip-gram encodings perform only slightly better than random and about on par with the speech rate selection method.

We used the submodular two-layer function with TF-IDF encodings (SM method) for subsequent investigations across all languages. Figures 1 and 2 show the active learning rates over selection points of 2-, 4-, 9-, and 14-hours, terminating once the whole selection pool has been added. The SM method generally performs better than the baseline heuristic method. The absolute gains vary across languages, with the best gains realized in Amharic. The addition of web data also tends to improve both the baseline and SM selections. Figure 4 provides a closer look at the gains offered by the SM method over the baseline method. In general, we see consistent gains up to the 14-hour point, though these vary by language.

We examine the effects of adding web data in further detail in Table 3, which show the drop in WER when web data is added to the SM selections and to the baseline selections. Augmenting the language model helps improve the performance of both methods across all languages except Igbo, which may be due to the relative scarcity of web data for Igbo (Table 2). Smaller gains are also realized with Pashto when compared to Amharic and Mongolian.

We also investigated adapting to the evaluation audio using the two-layer function (SM-adapt) with 2-, 4-, 9- and 14-hour selections on Mongolian. Table 4 shows changes in WER over the SM method. In general, we see gains at small selections with deterioration at the 14-hour point. We examined the effect of adding web data to SM-adapt selections and see that augmenting language models improves

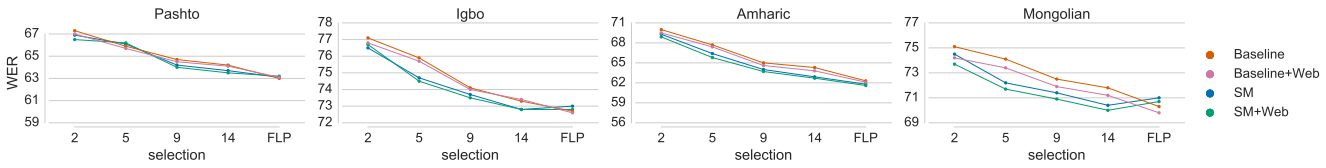

Fig. 1. *Active learning curves with context-dependent GMM based ASR.*

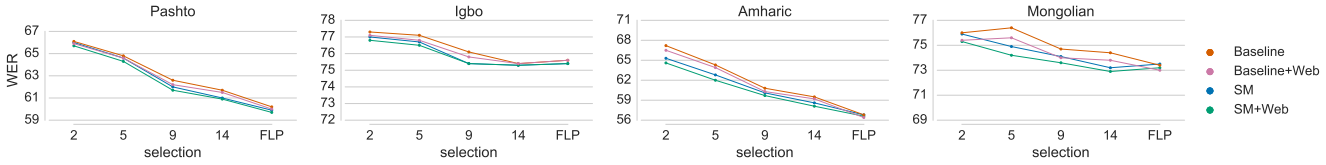

Fig. 2. *Active learning curves with DNN based ASR.*

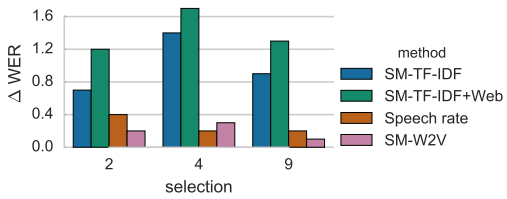

Fig. 3. *Mongolian: absolute improvements to WER over baseline at 2-, 4-, and 9-hours with different selection methods.*

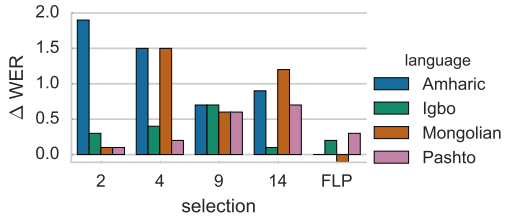

Fig. 4. *Submodular (TF-IDF) performance. Absolute improvements to WER over baseline at 2-, 4-, 9-, 14-hours and FLP.*

these selections, though gains are on par with the SM+Web method by the 14-hour point. Since SM-adapt seems most beneficial at smaller selections, we experimented with 2-hour selections across all languages. Table 4 shows that SM-adapt leads to small gains in WER over the SM method, while including web data leads to larger gains over the SM+Web method.

Language	Hrs	SM	SM+Web	Adapt	Adapt+Web
Mongolian	2	75.20	74.50	74.50	73.80
	4	73.55	72.95	73.35	72.75
	9	72.75	72.25	72.60	72.15
	14	71.80	71.45	72.00	71.50
Amharic	2	67.30	66.75	67.15	66.50
Pashto	2	66.45	66.10	66.35	66.10

Table 4. *Performance with adaptation to target audio.*

6. CONCLUSION

In the context of speech recognition on low-resource languages, we explored the value of active learning with larger selection sets, and with the introduction of more data for language modeling.

We find that, in general, the value of active learning is persistent with larger selections. We see larger gains from active learning over random selection at 4 hour selections than at 2 hours, with smaller but consistent gains at 9 and 14 hours selections. We assess the hypothesis that gains from active learning may be overwhelmed by the introduction of additional language model data and find performance improvements from active learning and additional language model data to be orthogonal and complementary. The impact of additional language data is fairly consistent within language, regardless of the amount of training data or selection strategy.

Automatic speech recognition comprises (at least) three interconnected problems: acoustic modeling, pronunciation modeling, and language modeling. The submodular functions we investigated for active learning address the acoustic modeling problem. They identify acoustic diversity and representativeness of each frame and aggregate over candidate utterances. This selection process does not explicitly identify diverse and representative productions (pronunciations) of given sequences, or word sequences. Any improvement to these criteria is only as a side effect of frame-based acoustic qualities. An optimal selection criteria for ASR would address all three problems. This remains a question for future work.

7. ACKNOWLEDGMENT

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD / ARL) contract number W911NF-12-C-0012. 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.

8. REFERENCES

[1] A Laurent, T Fraga-Silva, L Lamel, and J L Gauvain, "Investigating techniques for low resource conversational speech recognition," *ICASSP*, 2016.

[2] Thiago Fraga-Silva, Jean-Luc Gauvain, Lori Lamel, Antoine Laurent, Viet-bac Le, and Abdel Messaoudi, "Active Learning based data selection for limited resource STT and KWS," *INTERSPEECH*, 2015.

[3] Thiago Fraga-Silva, Antoine Laurent, Jean Luc Gauvain, Lori Lamel, Viet Bac Le, and Abdel Messaoudi, "Improving data selection for low-resource STT and KWS," *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015 - Proceedings*, 2016.

[4] C. Ni, C.-C. Leung, L. Wang, H. Liu, F. Rao, L. Lu, N.F. Chen, B. Ma, and H. Li, "Cross-lingual deep neural network based submodular unbiased data selection for low-resource keyword search," *ICASSP*, 2016.

[5] Ali Raza Syed, Andrew Rosenberg, and Ellen Kislal, "Supervised and Unsupervised Active Learning for Automatic Speech Recognition of Low-Resource Languages," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[6] Nancy F Chen, Chongjia Ni, I-fan Chen, Sunil Sivadas, Van Tung Pham, Haihua Xu, Xiong Xiao, Tze Siong Lau, Su Jun Leow, Boon Pang Lim, Cheung-chi Leung, Lei Wang, Chin-hui Lee, Alvina Goh, Eng Siong Chng, Bin Ma, and Haizhou Li, "Low-resource keyword search strategies for Tamil," in *ICASSP*, 2015.

[7] Giuseppe Riccardi and D. Hakkani-Tur, "Active learning: theory and applications to automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, 2005.

[8] Dilek Hakkani-Tur, Giuseppe Riccardi, and Allen Gorin, "Active learning for automatic speech recognition," in *ICASSP*, 2002.

[9] Kai Yu, Mark Gales, Lan Wang, and Philip C. Woodland, "Unsupervised training and directed manual transcription for LVCSR," *Speech Communication*, vol. 52, no. 7-8, pp. 652–663, 2010.

[10] Dong Yu, Balakrishnan Varadarajan, Li Deng, and Alex Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Computer Speech & Language*, , no. 3, 2010.

[11] Hui Lin and Jeff Bilmes, "How to Select a Good Training-data Subset for Transcription : Submodular Active Selection for Sequences," in *ICASSP*, 2009.

[12] Kai Wei, Yuzong Liu, Katrin Kirchhoff, Chris Bartels, and Jeff Bilmes, "Submodular subset selection for large-scale speech training data," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[13] Kai Wei, Yuzong Liu, Katrin Kirchhoff, and Jeff Bilmes, "Unsupervised submodular subset selection for speech data: Extended version," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. 2014, IEEE.

[14] Chongjia Ni, Cheung-chi Leung, Lei Wang, Nancy F Chen, and Bin Ma, "Unsupervised data selection and word-morph mixed language model for tamil low-resource keyword search," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, IEEE.

[15] Jack Edmonds, *Submodular Functions, Matroids, and Certain Polyhedra*, pp. 11–26, Springer Berlin Heidelberg, 2003.

[16] G. L. Nemhauser, L. a. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions-I," *Mathematical Programming*, vol. 14, no. 1, 1978.

[17] Andreas Krause and Daniel Golovin, "Submodular function maximization," *Tractability: Practical Approaches to Hard Problems*, vol. 3, 2014.

[18] Adarsh Prasad, Stefanie Jegelka, and Dhruv Batra, "Submodular Maximization and Diversity in Structured Output Spaces," in *NIPS*, 2014.

[19] Peter Stobbe and Andreas Krause, "Efficient Minimization of Decomposable Submodular Functions," *NIPS*, 2010.

[20] Hagen Soltau, George Saon, and Brian Kingsbury, "The IBM Attila speech recognition toolkit," in *2010 IEEE Spoken Language Technology Workshop*. 2010, IEEE.

[21] E. Khoury, L. El Shafey, and S. Marcel, "Spear: An open source toolbox for speaker recognition based on Bob," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[22] R. Villing, "Automatic blind syllable segmentation for continuous speech," 2004, vol. 2004, IEE.

[23] Andrew Rosenberg, "AuToBI A Tool for Automatic ToBI annotation," *Corpus*, , no. September, 2010.

[24] Zoltán Tüske, Pavel Golik, David Nolden, Ralf Schlüter, and Hermann Ney, "Data augmentation, feature combination, and multilingual neural networks to improve asr and kws performance for low-resource languages.," in *INTERSPEECH*. Citeseer, 2014.

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Distributed Representations of Words and Phrases and their Compositionality," *NIPS*, 2013.

[26] Gideon Mendels, Erica Cooper, Victor Soto, Julia Hirschberg, Mark Gales, Kate Knill, Anton Ragni, and Haipeng Wang, "Improving Speech Recognition and Keyword Search for Low Resource Languages Using Web Data," 2015.