CROSS-LINGUAL DEEP NEURAL NETWORK BASED SUBMODULAR UNBIASED DATA SELECTION FOR LOW-RESOURCE KEYWORD SEARCH

Chongjia Ni[1], Cheung-Chi Leung[1], Lei Wang[1], Haibo Liu[2], Feng Rao[2], Li Lu[2], Nancy F. Chen[1], Bin Ma[1], Haizhou Li[1]

[1] Institute for Infocomm Research (I^{2}R), A*STAR, Singapore

[2] Tencent Inc., Beijing, P. R. China

{nicj,ccleung,wangl,nfychen,mabin,hli}@i2r.a-star.edu.sg {geneliu,ralphrao,adolphlu}@tencent.com

ABSTRACT

In this paper, we propose a cross-lingual deep neural network (DNN) based submodular unbiased data selection approach for low-resource keyword search (KWS). A small amount (e.g. one hour) of transcribed data is used to conduct cross-lingual transfer. The frame-level senone sequence activated by the cross-lingual DNN is used to represent each untranscribed speech utterance. The proposed submodular function considers utterance length normalization and the feature distribution matched to a development set. Experiments are conducted by selecting 9 hours of Tamil speech for the 2014 NIST Open Keyword Search Evaluation (OpenKWS14). The proposed data selection approach provides 35.8% relative actual term weighted value (ATWV) improvement over random selection on the OpenKWS14 *Evalpart1* data set. Further analysis of the experimental results shows that both utterance length normalization and the feature distribution estimated from a development set deployed in the submodular function can suppress the preference to select long utterances. The selected utterances can cover a more diverse range of tri-phones, words, and acoustic variations from a wider set of utterances. Moreover, the wider coverage of words also benefits the acquired linguistic knowledge, which also contributes to improving KWS performance.

Index Terms— Submodular optimization, keyword spotting, spoken term detection, active learning

1. INTRODUCTION

The amount of multimedia data on the Internet has increased rapidly during the past decades, so the demand on efficient data indexing and search techniques is growing. Searching for keywords in spoken documents is still challenging because spoken documents are usually untranscribed [1-8]. Most state-of-the-art approaches for keyword search (KWS) are based on automatic speech recognition (ASR) [3-8]. With an ASR system, a lattice is generated for each spoken document in an archive. The lattices are converted to an inverted index, and KWS is performed on the inverted index.

Building an ASR system requires transcribed speech and linguistic knowledge. However, it is both time-consuming and costly to manually transcribe the speech data as well as to construct a lexicon for a particular language, especially for a low-resource language. This motivates us to build an efficient ASR system using as little transcribed data as possible. To achieve this, one way is to select representative speech utterances for manual transcription instead of transcribing the entire set.

In this work, to accomplish the KWS task of a low-resource language, we assume that an initial small amount (e.g. one hour) of manually transcribed speech data and a large amount of untranscribed speech data are available. The data selection problem is to select a subset of the untranscribed data for manual transcription. The words that we find from the manual transcription are included to construct a pronunciation dictionary. The selected data together with the initially transcribed data is then used to build an ASR system for KWS. In this paper, we would also like to study which factors are important for improving KWS performance.

A preliminary investigation of Gaussian component index based submodular data selection for low-resource keyword search has been reported in our previous work [9]. It is an unsupervised data selection approach. This work further extends [9] in two aspects. Firstly, we use the initially transcribed data as a seed to select untranscribed data for manual transcription. We use the small amount of data to conduct cross-lingual transfer, then the frame-level senone sequence activated by the cross-lingual DNN is used to represent each untranscribed utterance, and then the submodular based data selection approach is used to select the untranscribed utterances for manual transcription; Secondly, we propose a novel objective function for submodular optimization. In the objective function, utterance length normalization and the feature distribution matched to a development set are both considered to further improve KWS performance. Utterance length normalization is aimed to increase speaker diversity by suppressing the preference to select long utterances.

The effect of data selection on KWS performance, to the best of our knowledge, is not well studied. The most related study is our previous work [9], in which using Gaussian component index based n-grams as features in the submodular function does not require an initial ASR system, and the submodular function provides a near-optimal solution in terms of the objective being optimized. However, it is worth noting that different kinds of active learning techniques have been investigated to address the data selection problem for phone recognition and ASR. For unsupervised submodular based data selection, a Fisher-kernel based graph [10] over untranscribed utterances was proposed, but this approach requires computing the similarity between all utterance pairs. In a later work, a two-level feature-based submodular function was proposed to select a subset of untranscribed data in the TIMIT corpus for training phone recognizers [11].

Another type of related work is semi-supervised and supervised data selection. Confidence-based approaches [12-17] are commonly used for semi-supervised data selection, in which an initial ASR system is available to obtain phone hypotheses of the untranscribed data. The informative utterances are selected and

used to update the initial ASR system. Itoh *et al.* [18] suggested that both informativeness and representativeness of the data should be assessed. Siohan *et al.* [19, 20] proposed to use i-vector and senone sequences to represent utterances, and use relative-entropy data selection algorithm to select utterances, so the acoustic or phonetic distribution of the informative utterances was considered. Wu *et al.* proposed to choose data uniformly according to the distribution of the target speech units [21]. However, these approaches cannot provide any optimality guarantee as those based on submodular optimization.

Submodular optimization was also considered in semi-supervised and supervised data selection. Wei *et al.* proposed the feature-based submodular data selection approach using phone n-grams as features [22] to select a subset of labeled data to build an ASR system. Wei *et al.* proposed to use the string kernel submodular data selection [23] for phone recognition, which was based on the hypothesis of each utterance derived by a phone recognizer. Shinohara considered a desired (uniform) phone distribution in the submodular function [24] for speech data selection.

2. SUBMODULAR FUNCTION AND OPTIMIZATION

A set-valued function $\boldsymbol{f}: \boldsymbol{2^V} \rightarrow \mathbb{R}$ is a submodular function if for every subset $\mathbf{A}$ and $\mathbf{B}$ in a finite set $\boldsymbol{V} = \{\mathbf{1}, \mathbf{2}, \cdots, \boldsymbol{N}\}$ with $\mathbf{A} \subseteq \mathbf{B}$ and each item $\boldsymbol{s} \in \boldsymbol{V} \backslash \boldsymbol{B}$,

$$f(B \cup \{s\}) - f(B) \leq f(A \cup \{s\}) - f(A). \quad (1)$$

Submodularity is a property of set-valued function, and it means that the gain by adding an element into a smaller set should not be less than that by adding the element into a superset. Different submodular functions have been proposed and used in speech data selection [9-11, 22-25].

A submodular function f is monotone non-decreasing if $f(A \cup \{s\}) - f(A) \geq 0$ for $\forall s \in V \backslash A, A \subseteq V$. A submodular function f is normalized if $f(\emptyset) = 0$, and Ø is empty set.

The subset selection problem can be formulated as follows:

$$\max_{S \subseteq V} \{f(S): c(s) \leq K\} \quad (2)$$

where $\boldsymbol{c}(\boldsymbol{s}) \leq \boldsymbol{K}$ is a constraint. For ASR application, a subset $\boldsymbol{S}$ of training data is selected from $\boldsymbol{V}$ by maximizing the objective function $\boldsymbol{f}$ at the constraint $\boldsymbol{c}(\boldsymbol{s}) \leq \boldsymbol{K}$. The constraint can be the number of selected utterances, or the number of hours of the selected utterances.

The subset selection problem is NP hard. But it can be approximately solved using a simple greedy forward selection algorithm. The solution is guaranteed to be near-optimal [26], and it is the best we can do in polynomial time unless $P = NP$[27].

3. DATA SELECTION FOR ACOUSTIC MODELING

3.1. Cross-lingual DNN based utterance representation

Cross-lingual knowledge transfer is an efficient approach to improve the performance of low-resource ASR by transferring the knowledge of rich-resource language(s) to a low-resource language. Borrowing the feature extractor or acoustic model of rich-resource language(s) at deep neural network framework is common way for knowledge transfer in ASR [28-33]. The shared-hidden-layer multilingual deep neural network (SHL-MDNN) framework for cross-lingual knowledge transfer has been successfully applied in low-resource ASR or KWS [33]. At the SHL-MDNN framework, different languages share hidden layers except the language-dependent softmax layer.

It is difficult to build an efficient ASR system with the initial small amount of transcribed data. An efficient approach is to conduct cross-lingual transfer using the small amount of target-language data. We stack the softmax layer of the target language to the shared hidden layers of the SHL-MDNN, and only the softmax layer is updated when the small initial amount of target-language training data is presented to the model.

Each untranscribed utterance is decoded by the cross-lingual DNN to generate a frame-level senone sequence. We count the n-grams of the senone sequence, and the term frequency-inverse document frequency (tf-idf) of the n-grams is used to represent each utterance in a fixed-dimensional vector.

3.2. Feature based submodular function

There are different submodular functions proposed for data selection. The feature-based submodular function $f_{fea}(S) = \sum_{u \in U} g\big(m_u(S)\big)$ is commonly adopted for data selection, where $m_u(S) = \sum_{s \in S} m_u(s)$ measures the degree of feature u in the subset S [22], $m_u(s)$ is defined by the tf-idf vector of utterance s, and $g(\cdot)$ is a monotone non-decreasing function. Compared with the facility location based submodular function in [10], the feature-based submodular function $f_{fea}(\cdot)$ does not need to compute the similarity between two utterances, so it can obtain lower computation complexity.

A shortcoming of the feature-based submodular function $f_{fea}(\cdot)$ is that it prefers to select long utterances. If the constraint is the total number of hours of the selected utterances and the number of selected utterances is small, the selected long utterances could limit the overall acoustic variation (such as speaker diversity). In [9], the development set matching based submodular function $f_{dev-matched-fea}(S) = \sum_{u \in U} p_u log(m_u(S))$ is used for data selection, where $\{p_u\}$ is the distribution of feature $u \in U$, and is estimated from a development set. Using $\{p_u\}$ as feature weights is aimed to target a subset of features in the selection. Empirically we find that this alleviates the preference of selecting long utterances.

In this paper, we would further reduce the preference of selecting long utterances. We normalize the measurement of the degree of feature u in utterance s by its utterance length. The proposed function is as follows:

$$\begin{aligned} f_{dev-matched-fea+len-norm}(S) &= \sum\nolimits_{u \in U} p_u \log\Big(\sum\nolimits_{s \in S} \frac{1}{l(s)} m_u(s)\Big) \\ &= \sum\nolimits_{u \in U} p_u \log\big(m_u^*(S)\big) \end{aligned} \quad (3)$$

where $l(s)$ is the length of utterance s, $m_u^*(S) = \sum_{s \in S} \frac{1}{l(s)} m_u(s)$ measures the average degree of feature u in the subset S. Eq. (3) is a submodular function according to submodular theory [26, 27, 34].

4. EXPERIMENTS

4.1. Experimental setup

The Tamil data provided by the IARPA Babel program for OpenKWS14 was used in experiments. There are two conditions which correspond to two sets of training data. One is full language pack (FLP), and another is limited language pack (LLP). In FLP

condition, all data resources including 60 hours of transcribed audio can be used to build a keyword search system. LLP contains a subset of 10 hours of transcribed audio in FLP, and the remaining data in LLP is untranscribed audio from the remaining data in FLP. The audio data is conversational telephone speech obtained from different channels, like landlines, cell phones, and phones embedded in vehicles. The scripted speech which is aimed to improve the coverage of phonemes is also included in FLP. The speech from four languages (including Cantonese, Pashto, Turkish, and Tagalog) in the Babel corpora was used for multilingual DNN training. For the four languages, more than 100 hours of data are recorded in each full language pack (FLP), and pronunciation lexicon only covers the word appeared in training transcription. One hour of transcribed data used for cross-lingual knowledge transfer was randomly selected from FLP.

For evaluating our proposed approach, we built three baseline systems. The first baseline system (denoted as LLP) was built using the training data in LLP. The second baseline system (denoted as FLP-10h-Random) was built using randomly selected 10 hours of data from FLP, and its performance was reported by an average result of two runs. The third baseline system (denoted as MaxEnt) was built using the maximum entropy approach proposed in [21] to select 10 hours of data from the 60-hour transcribed audio in FLP. We also built a topline system (denoted as FLP) using the 60 hours of training data in FLP.

All KWS systems are word based. The actual term weighted value (ATWV) and word error rate (WER) were used to measure the performance of keyword search systems and the underlying ASR systems. The 10 hours of development set *Dev10h* and 15 hours of evaluation part 1 *Evalpart1* were used for our evaluation. The keyword list provided for OpenKWS14 which contains 5,576 keywords or keyword phrases was used for evaluating keyword search systems. All hybrid DNN acoustic models were built as in our previous work [9]. The tri-gram language models trained using corresponding training transcriptions were used for lattice generation. The number of activated non-silence senones was used to measure the length of utterance in Eq. (3). The Kaldi toolkit and its KWS recipe [35] were used in our experiments.

4.2. Experimental results

Table 1 lists the performance of different keyword search systems built using different sets of data. "FLP-Proposed" was built using our proposed approach, in which one hour of data was initially available and 9 hours of data was selected from the remaining 59 hours of audio in FLP. From Table 1, we can see that: (1) The WER is high for all underlying ASR systems due to the low-resource condition (in both acoustic models and language models); (2) The performance of the keyword search system built using our proposed approach is better than that of the three baseline systems, and there are 27.8%, 35.8% and 42% relative ATWV improvements on *Evalpart1* when comparing with "LLP", "FLP-10h-Random" and "MaxEnt" respectively.

Table 1. Comparison of different data selection approaches

	Data Set	FLP	LLP	FLP-10h-Random	MaxEnt	FLP-Proposed
WER (%)	*Dev10h*	64.4	75.4	78.7	78.9	71.8
	Evalpart1	66.1	77.0	79.8	79.9	71.6
ATWV	*Dev10h*	0.4349	0.2336	0.2203	0.2096	0.3028
	Evalpart1	0.4222	0.2313	0.2199	0.2081	0.2986

In order to analyze the effect of different utterance representation approaches for data selection, we selected $f_{dev-matched-fea+len-norm}(\cdot)$ as the submodular function, and conducted experiments using different utterance representations: senone sequences decoded by a cross-lingual DNN, Gaussian component index sequences as in [9], and phone sequences decoded by the BUT Hungarian phone recognizer [36]. Table 2 lists the experimental results. We believe that the cross-lingual DNN and the Gaussian mixture model, which suffer less from acoustic data mismatch, can provide more accurate utterance representations and select more representative utterances for building ASR and KWS systems.

Table 2. Comparison of different utterance representations

	Data Set	Cross-lingual DNN Senones	Gaussian Component Indices	Phones from BUT Hungarian recognizer
WER (%)	*Dev10h*	71.8	71.6	77.3
	Evalpart1	71.6	73.3	78.4
ATWV	*Dev10h*	0.3028	0.2982	0.1874
	Evalpart1	0.2986	0.2916	0.1991

Table 3. Comparison of different submodular functions

	WER (%)		ATWV	
	Dev10h	Evalpart1	Dev10h	Evalpart1
$f_{fea}(\cdot)$	80.6	81.3	0.1249	0.1271
$f_{dev-mateched-fea}(\cdot)$	73.3	74.4	0.2952	0.2850
$f_{dev-matched-fea+len-norm}(\cdot)$	71.6	73.3	0.2982	0.2916

In order to analyze the effect of different submodular functions for KWS, we conducted experiments using different submodular functions with the Gaussian component index based utterance representation. Table 3 lists the experimental results. From Table 3, we observe that: (1) $f_{dev-matched-fea}(\cdot)$ which includes $\{p_u\}$ estimated from the development set as feature weights in the submodular function, obviously outperforms $f_{fea}(\cdot)$; (2) When comparing with $f_{dev-matched-fea}(\cdot)$ used in our previous work [9], $f_{dev-matched-fea+len-norm}(\cdot)$ provides 4.8% relative ATWV improvement on *Evalpart1*.

In addition to WERs and ATWVs, we further investigated the difference of the selected utterances between using the three submodular functions. In each set of selected data, together with the corresponding training transcriptions, we measured the average length of the utterances (denoted as "Avg. Length"; in second), the total number of unique cross-word tri-phones (denoted as "#Triphone"), the average number of occurrences of each tri-phone (denoted as "Avg. Occ."), the total number of unique words (denoted as "|L|"), and the number of unique out-of-vocabulary keywords or keyword phrases (denoted as #OOV). The statistical results are summarized in Table 4.

Table 4. Statistics of the selected utterances using different submodular functions

	Avg. Length	#Tri-phone	Avg. Occ.	$\|L\|$	#OOV
$f_{fea}(\cdot)$	11.17	9725	25.1	11984	1910
$f_{dev-mateched-fea}(\cdot)$	8.02	11669	45.0	18844	972
$f_{dev-matched-fea+len-norm}(\cdot)$	1.72	11719	47.8	18694	1309

From the statistical results, we observe that: (1) $f_{fea}(\cdot)$ prefers to select longer utterances. It covers fewer tri-phones and fewer

words, and leads to more out-of-vocabulary keywords or keyword phrases. These explain why it provides the highest WER and the poorest ATWV; (2) By considering the feature distribution estimated from the development set in the submodular function, it can cover more tri-phones and more words, and reduce the number of out-of-vocabulary keywords or keyword phrases. These lead to better keyword search performance; (3) When utterance length normalization is considered in the submodular function, the preference to select longer utterances is further suppressed. This leads to further cover more tri-phones, more utterances, and more acoustic variations from different utterances. And this is probably the reason why $f_{dev-matched-fea+len-norm}(\cdot)$ performs the best in terms of WER and ATWV.

The number of words in the lexicon given in the LLP is 14,269. When our proposed data selection approach was used, the number of words in the lexicon increased to 18,694. Since our proposed approach can select more words, and we believe that the linguistic knowledge (including lexicon and language model) acquired from the corresponding training transcription is more information rich. To further study how the acoustic model and the linguistic knowledge acquired by our proposed approach affect the keyword search performance, we utilized different lexicons (with LMs updated according to their lexicons) for keyword search with the same acoustic model of the keyword search system "FLP-Proposed". Table 5 lists the experimental results.

From Table 5, we can find that: (1) With the lexicon of FLP, the keyword search system can obtain better performance though it performs worse than the topline system (FLP) reported in Table 1; (2) With the lexicon of LLP, the keyword search system performs worse comparing with our proposed system. The acoustic model trained using our selected data is better than the acoustic model trained using the data in LLP (see the LLP column in Table 1); (3) While the linguistic knowledge acquired by our proposed approach helps improving the performance of KWS and the underlying ASR system, the ATWV is more sensitive to the acquired linguistic knowledge than the WER.

Table 5. KWS Performance using different lexicons. All systems use the acoustic model "FLP-Proposed".

	WER (%)		ATWV	
	Dev10h	**Evalpart1**	**Dev10h**	**Evalpart1**
Proposed Lexicon	71.8	71.6	0.3028	0.2986
LLP Lexicon	72.2	73.3	0.2927	0.2726
FLP Lexicon	70.2	71.3	0.3451	0.3491

As manually constructing a lexicon is time-consuming and costly, one potential solution of building a keyword search system is to ignore the out-of-vocabulary words by modeling them using a garbage model. The keyword search system is denoted as "Fixed-LLP-Lex" in Table 6.

The keyword search performance of "Fixed-LLP-Lex" dropped by at least 0.0737 in terms of ATWV on *Evalpart1*, when comparing with "FLP-Proposed" in which a manual lexicon was used. Many words were modeled using the garbage model so that the acoustic models could not be well trained, which led to poor keyword search performance. We also built a keyword search system, which used the phonetisaurus grapheme-to-phoneme (G2P) toolkit [37] to automatically acquire the pronunciation of the new words discovered by our proposed approach. This keyword search system is denoted as "LLP-Lex-Plus-G2P" and its performance is also listed in Table 6. "LLP-Lex-Plus-G2P" performed better than "Fixed-LLP-Lex", and performed comparable with "FLP-Proposed" in which a manual lexicon was used.

Table 6. KWS performance using different lexicon conditions. All systems use acoustic model "FLP-Proposed".

	WER (%)		ATWV	
	Dev10h	**Evalpart1**	**Dev10h**	**Evalpart1**
Manual lexicon	71.8	71.6	0.3028	0.2986
Fixed-LLP-Lex	74.5	75.9	0.2368	0.2249
LLP-Lex-Plus-G2P	71.9	72.8	0.3028	0.2881

5. CONCLUSIONS

In this paper, we propose to use a frame-level senone sequence decoded by a cross-lingual DNN to represent each untranscribed utterance, and use the submodular function, which consider utterance length normalization and the feature distribution matched a development set, to select utterances for transcription for further improving the performance of a low-resource keyword search system. Experiment results show that n-grams of senone sequences provide a kind of utterance representation with performance comparable to that provided by n-grams of Gaussian component indices. And it is shown that both utterance length normalization and the feature distribution estimated from a development set deployed in the submodular function can suppress the preference to select long utterances. This can lead to the selected utterances to cover more different tri-phones and words, and more acoustic variations from different utterances. If the lexicon needs to cover all the words discovered in the selected utterances, this unavoidably increases the budget for updating the lexicon. Our experiment shows that using G2P to automatically acquire the pronunciation of the new words can accomplish a comparable ATWV as manually updating the lexicon. This strategy is especially practical for a low-resource setting.

6. REFERENCES

[1] J. G. Wilpon, L. R. Rabiner, C.-H. Lee and E. Goldman, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1990, 38(11), 1870-1878.

[2] R. C. Rose and D. B. Paul, "A Hidden Markov Model based Keyword Recognition System," in *Proc. ICASSP* 1990, pp. 129-132.

[3] J. Mamou, B. Ramabhadran, and O. Siohan, "Vocabulary Independent Spoken Term Detection," in *Proc. SIGIR 2007*, pp. 615-622.

[4] D. R. H. Miller, M. Kleber, C.-L. Kao, O. Kimball, T. Colthurst, and S. A. Lowe, "Rapid and Accurate Spoken Term Detection," in *Proc. Interspeech 2007*.

[5] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddintion, "Results of the 2006 Spoken Term Detection Evaluation," in *Proc. SIGIR* 2007.

[6] I. Szoeke, M. Fapso, and L. Burget, "Hybrid Word-subword Decoding for Spoken Term Detection," in *Proc. SIGIR* 2008.

[7] N. F. Chen, S. Sivadas, B. P. Lim, H. G. Ngo, H. Xu, V. T. Pham, B. Ma, H. Li, "Strategies for Vietnamese Keyword Search," in *Proc. ICASSP* 2014, pp.4121-4125.

[8] N. F. Chen, C. Ni, I-F. Chen, S. Sivadas, V. T. Pham, H. Xu, X. Xiao, T. S. Lau, S. J. Leow, B. P. Lim, C.-C. Leung, L.

Wang, C.-H. Lee, A. Goh, E. S. Chng, B. Ma, H. Li, "Low-Resource Keyword Search Strategies for Tami," in *Proc. ICASSP* 2015.

[9] C. Ni, C.-C. Leung, L. Wang, N. F. Chen and B. Ma, "Unsupervised Data Selection and Word Morph Mixed Language Model for Tamil Low Resource Spoken Keyword Spotting," in *Proc. ICASSP* 2015.

[10] H. Lin and J. Bilmes, "How to Select a Good Training-data Subset for Transcription: Submodular Active Selection for Sequences," in *Proc. Interspeech 2009*, pp. 2859-2862.

[11] K. Wei, Y. Liu, K. Kirchhoff, and J. Bilmes, "Unsupervised Submodular Subset Selection for Speech Data," in *Proc. ICASSP 2014*, pp. 4107-4111.

[12] Y. Huang, D. Yu, Y. Gong, and C. Liu, "Semi-Supervised GMM and DNN Acoustic Model Training with Multi-system Combination and Confidence Re-calibration," in *Proc. Interspeech 2013*, pp. 2360-2364.

[13] K. Vesely, M. Hannemann, L. Burget, "Semi-Supervised Training of Deep Neural Networks," in *Proc. ASRU 2013*, pp. 267-272.

[14] D. Charlet, "Confidence-measure-driven Unsupervised Incremental Adaptation for HMM-based Speech Recognition," in *Proc. ICASSP 2001*, pp. 357-360.

[15] G. Tur, D. Hakkani-Tur and R. E. Shapire, "Combining Active and Semi-supervised Learning for Spoken Language Understanding," *Speech Communication*, 2005, 45(2):171-186.

[16] X. Zhu, "Semi-supervised Learning Literature Survey," *Computer Sciences Technical Report 1530*, University of Wisconsin-Madison, 2005b.

[17] D. Yu, B. Varadarajan, L. Deng and A. Acero, "Active Learning and Semi-Supervised Learning for Speech Recognition: A Unified Framework Using the Global Entropy Reduction Maximization Criterion," *Computer Speech and Language*, 2010, 24(3): 433-444.

[18] N. Itoh, T. N. Sainath, D. N. Jiang, J. Zhou, and B. Ramabhadran, "N-Best Entropy Based Data Selection for Acoustic Modeling," in *Proc. ICASSP 2012*, pp. 4133-4136.

[19] O. Siohan, and M. Bacchiani, "iVector-based Acoustic Data Selection," in *Proc. Interspeech 2013*, pp. 657-661.

[20] O. Siohan, "Training Data Selection Based on Context-Dependent State Matching," in *Proc. ICASSP 2014*, pp. 3316-3319.

[21] Y. Wu, R. Zhang, and A. Rudnicky, "Data Selection for Speech Recognition," in *Proc. ASRU 2007*, pp.562-565.

[22] K. Wei, Y. Liu, K. Kirchhoff, C. Bartels and J. Bilmes, "Submodular Subset Selection for Large-Scale Speech Training Data," in *Proc. ICASSP 2014*, pp. 3311- 3315.

[23] K. Wei, Y. Liu, K. Kirchhoff, and J. Bilmes, "Using Document Summarization Techniques for Speech Data Subset Selection," in *Proc. NAACL/HLT-2013*, pp. 721-726.

[24] Y. Shinohara, "A Submodular Optimization Approach to Sentence Set Selection," in *Proc. ICASSP 2014*, pp. 4140-4143.

[25] C. Ni, L. Wang, H. Liu, C.-C. Leung, L. Lu, and B. Ma, "Submodular Data Selection with Acoustic and Phonetic Features for Automatic Speech Recognition," in *Proc. ICASSP 2015*.

[26] G. Nemhauser, L. Wolsey, and M. Fisher, "An Analysis of Approximations for Maximizing Submodular Set Function-I," *Mathematical Programming*, 1978, 14(1):265-294.

[27] U. Feige, "A Threshold of ln n for Approximating Set Cover," *Journal of the ACM*, 1998, 45(4):634-652.

[28] Y. Zhang, E. Chuangsuwanich, J. Glass, "Language ID-based Training of Multilingual Stacked Bottleneck Features," in *Proc. Interspeech 2014*, pp.1-5.

[29] K. Vesely, M. Karafiat, F. Grezl, M. Janda, and E. Egorova, "The Language-independent Bottleneck Features," in *Proc. SLT 2012*, pp.336-340.

[30] K. M. Knill, M. J. F. Gales, S. P. Rath, P. C. Woodland, C. Zhang, and S.-X. Zhang, "Investigation of Multilingual Deep Neural Networks for Spoken Term Detection," in *Proc. ASRU 2013*.

[31] Z. Tuske, D. Nolden, R. Schluter, H. Ney, "Multilingual MRASTA Features for Low-resource Keyword Search and Speech Recognition Systems," in *Proc. ICASSP 2014*, pp.7854-7858.

[32] A. Ghoshal, P. Swietojanski, S. Renals, "Multilingual Training of Deep Neural Networks," in *Proc. ICASSP 2013*.

[33] J.-T Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language Knowledge Transfer using Multilingual Deep Neural Network with Shared Hidden Layers," in *Proc. ICASSP 2013*, pp. 7304-7308.

[34] F. Bach, "Learning with Submodular Functions: A Convex Optimization Perspective," *Foundations and Trends ® in Machine Learning*," 2013, 6(2-3): 145-373.

[35] 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*, 2011.

[36] P. Schwarz, "Phoneme Recognition based on Long Temporal Context, PhD Thesis," Brno University of Technology, 2009.

[37] phonetisaurus - A WFST-driven Phoneticizer, Available Online: https://code.google.com/p/phonetisaurus/.