INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Rapid Update of Multilingual Deep Neural Network for Low-Resource Keyword Search

Chongjia Ni[1], Lei Wang[1], Cheung-Chi Leung[1], Feng Rao[2], Li Lu[2], Bin Ma[1], and Haizhou Li[1]
[1] Institute for Infocomm Research (I²R), A*STAR, Singapore
[2] Tencent Inc., Beijing, P. R. China
{nicj,wangl,ccleung,mabin,hli}@i2r.a-star.edu.sg {ralphrao,adolphlu}@tencent.com

Abstract

This paper proposes an approach to rapidly update a multilingual deep neural network (DNN) acoustic model for low-resource keyword search (KWS). We use submodular data selection to select a small amount of multilingual data which covers diverse acoustic conditions and is acoustically close to a low-resource target language. The selected multilingual data together with a small amount of the target language data are then used to rapidly update the readily available multilingual DNN. Moreover, the weighted cross-entropy criterion is applied to update the multilingual DNN to obtain the acoustic model for the target language. To verify the proposed approach, experiments were conducted based on four speech corpora (including Cantonese, Pashto, Turkish, and Tagalog) provided by the IARPA Babel program and the OpenKWS14 Tamil corpus. The 3-hour very limited language pack (VLLP) of the Tamil corpus is considered as the target language, while the other four speech corpora are viewed as multilingual sources. Comparing with the traditional cross-lingual transfer approach, the proposed approach achieved a 19% relative improvement in actual term weighted value on the 15-hour evaluation set in the VLLP condition, when a word-based or word-morph mixed language model was used. Furthermore, the proposed approach was observed to have similar performance as the KWS system based on the acoustic model built using the target language and all multilingual data from scratch, but with shorter training time.

Index Terms: cross-lingual knowledge transfer, multilingual deep neural network, keyword spotting, very limited language pack condition, submodular data selection

1. Introduction

Low-resource keyword search (KWS) has become a focus area for several research groups over the past years. The traditional ways to search for a keyword in a speech corpus include keyword-filler based approaches [1-2] and large vocabulary continuous speech recognition (LVCSR) based approaches [3-8]. In this paper, we are interested in the LVCSR based approaches because they scale better when the number of search keywords increases.

As an LVCSR based KWS system needs a well-trained ASR system, it is difficult to build such a system for a low-resource language due to the insufficient quantity of transcribed speech data. To overcome such limitation, researchers proposed to use the transcribed data from other languages to build acoustic models or feature extractors which can then be applied to low-resource target languages [9-19]. To achieve this goal, cross-language knowledge transfer using multilingual deep neural networks (DNNs) with different training strategies [9,11,13,14,17,18] have been proposed. Another commonly used approach [10,12,15,16,19] is to extract bottleneck features using a multilingual DNN, which is trained using a large amount of transcribed data, and then train a GMM-HMM recognizer using the bottleneck features. A multilingual DNN trained by multiple source languages for bottleneck feature extraction or cross-lingual transfer carries rich information to distinguish phonetic classes in multiple source languages, and it can be adopted to distinguish phones in other target languages. Multilingual DNN based cross-lingual transfer can be simple and time-efficient when a small amount of transcribed target language data is available. Its shared hidden layers remain unchanged and only the softmax layer is trained using the target language data [14].

In the present work, we demonstrate that although the model obtained by cross-lingual transfer is time-efficient, the acoustic model trained using both the target and source language data outperforms the model obtained by cross-lingual transfer. We believe that the acoustic model that include both types of training data can capture more target language related phonetic information, especially when target language is from a language family different from the source languages. Unfortunately, the above mentioned approach has two issues: i) It takes a longer time to estimate the parameters of the DNN due to the large amount of multilingual training data and the large-scale of parameters in the DNN; ii) Not all multilingual data contributes equally to the final KWS performance for the target language [9].

Motivated by the two issues, this work examines an efficient way to update the multilingual DNN for the target language, instead of training the multilingual DNN from scratch. We assume that the multilingual DNN is readily available before a new target language is identified. It is desirable to rapidly deploy a KWS system for the new target language when the amount of transcribed target language data is limited.

To update the multilingual DNN for a new target language, we propose to use submodular data selection to select a subset of representative utterances from the multilingual data corpus. A selection metric is used to choose the utterances which are acoustically close to the target language. Representative utterances are defined as the utterances that cover diverse acoustic conditions. Submodular data selection has been shown successful in active learning for ASR and KWS [21-25], in which representative utterances from a language are selected for manual transcription. Note that multilingual data selection has been studied in [9] which used language identification to select the most similar language as the target language. However, our proposed approach considers data representativeness and it can select utterances from different language corpora. Furthermore, to

http://dx.doi.org/10.21437/Interspeech.2016-53

well re-estimate the parameters in the multilingual DNN model, we apply the weighed cross-entropy criterion to enhance the feedback error observed from the target language training data.

To examine our proposed approach, the above acoustic model is evaluated on KWS using the Tamil speech in OpenKWS14. Because the rich morphological structure of Tamil, both word-based and word-morph mixed language models (LMs) are used in the KWS experiments. To our best knowledge, it is the first work to use submodular data selection to select multilingual data to conduct multilingual DNN update. The rest of the paper is organized as follows. The background of multilingual DNN training is introduced in Section 2. The shared-hidden-layer multilingual deep neural network (SHL-MDNN) update with a new target language is presented in Section 3. Multilingual data selection for SHL-MDNN is introduced in Section 4. Experimental setup and results are presented in details in Section 5. Section 6 concludes the paper.

2. Background of multilingual DNN training

There are several ways to train a multilingual DNN that uses a universal or language dependent phone set [9-19, 26]. The shared-hidden-layer multilingual deep neural network (SHL-MDNN) [14] is one of the widely used approaches, in which the hidden layers are shared across many languages while the softmax layers are language dependent. The shared hidden layers (SHLs) extracted from the multilingual deep neural network can be viewed as a universal feature extraction module. As SHLs are trained using multiple source languages, they carry information for phonetic classification in the multiple source languages. When conducting cross-lingual transfer, the SHLs are extracted from the SHL-MDNN, and a new softmax layer is added on top of the SHLs. The output nodes in new softmax layer correspond to the senones created for the target language. Fixing the hidden layers, only the softmax layer is trained using the limited target language data. If sufficient training data is available, the entire network can be re-estimated for additional gains.

3. SHL-MDNN update with new target language

When adding a new language into the SHL-MDNN training for combining more rich phonetic information into modeling process, a simple approach is to train a SHL-MDNN from scratch using the new target language data and previous multilingual training data. It is undesirable to train a new SHL-MDNN from scratch due to the following two reasons: (1) The existing SHL-MDNN model has embodied the previously available multilingual training data; and (2) It will take a long time to train a new SHL-MDNN due to the large amount of multilingual training data.

The objective of combining the target language data and source language data is to build a better DNN model for the target language. In this paper, we propose to use the weighted cross-entropy criterion and a small amount of multilingual data to achieve this goal. We refer to this process as the SHL-MDNN update with a new target language. The weighted cross-entropy criterion can be used to emphasize more on the new target language when conducting SHL-MDNN update.

3.1. Weighted cross-entropy criterion for shared-hidden-layer multilingual deep neural network

The weighted cross-entropy criterion [27-29] used for SHL-MDNN training can be formulated as follows:

$$E = (1-\alpha)\sum_{j=1}^{M}\tilde{d}_j \log\tilde{p}_j + \frac{\alpha}{C}\sum_{k=1}^{C}\sum_{j=1}^{N_k} d_j^{(k)} \log p_j^{(k)} \quad (1)$$

where $\tilde{p}_j = \frac{\exp(\tilde{x}_j)}{\sum_k \exp(\tilde{x}_k)}$ is the probability of the target language output unit $\tilde{o}_j$ for its input $\tilde{x}_j$. $p_j^{(i)} = \frac{\exp(x_j^{(i)})}{\sum_k \exp(x_k^{(i)})}$ is the probability of the i^{th} source language output unit $o_j^{(i)}$ for its input $x_j^{(i)}$. $\tilde{d}_j$ and $d_j^{(k)}$ are target values, which are 1 when the input belongs to the j^{th} state and are 0 otherwise. C is the number of source languages. M is the number of senones in the target language, and N_k is the number of senones of the k^{th} source language. α is the weight. A higher target language related weight $(1-\alpha)$ can be used to emphasize more on the target language (and emphasize less on the source languages).

4. Multilingual data selection for SHL-MDNN update

Comparing with training an SHL-MDNN from scratch using all the speech data, it is more efficient to update the readily available SHL-MDNN using a small amount of selected multilingual training data and the training data of the new target language. The previous work [9] showed that not all multilingual data can contribute to modeling parameter estimation for the target language. Therefore, when selecting a small amount of multilingual training data for updating the SHL-MDNN, the selected data should not harm the target language related DNN model. In this paper, we propose to use submodular data selection to select a small amount of multilingual data, which is acoustically close to the target language data.

4.1. Problem formulation

Given a set of N utterances $V = \{v_1, v_2, \cdots, v_N\}$, $f: 2^V \rightarrow \mathbb{R}$, returning a real value for any subset $S \subseteq V$, is a submodular function if it satisfies $f(B \cup \{s\}) - f(B) \leq f(A \cup \{s\}) - f(A)$ $\forall A, B \subseteq V, A \subseteq B, s \in V \backslash B$. A function f is monotone non-decreasing if $f(A \cup \{s\}) - f(A) \geq 0$, $\forall s \in V \backslash A, A \subseteq V$. A function f is normalized if $f(\emptyset) = 0$.

For submodular subset selection problem, it can be formulated as follows:

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

where $c(S) \leq K$ is the constraint. In this paper, $c(S)$ is the number of hours of the utterances selected from the multilingual sources. Although the subset selection problem is NP hard, it can be approximately solved using a simple greedy forward selection algorithm. And the theorem in [30] guarantees that the solution obtained using the greedy forward selection algorithm is near-optimal. It is also the best we can do in polynomial time unless $P = NP$ [31].

4.2. Data selection using feature based submodular function

When conducting speech data selection for manual transcription, the feature based submodular data selection has been shown more efficient than other submodular data

selection [21-25], and to perform better results than other data selection approaches for ASR and KWS applications [23-25]. It does not require compute the pair-wise similarities between all the utterances when selecting utterances to form a suboptimal subset under the objective function optimization framework.

In this paper, we follow the submodular function $f_{\mathrm{dev-match+len-norm}}(S) = \sum_{u\in U} p_u \log\left(\sum_{s\in S}\frac{1}{l(s)} m_u(s)\right) = \sum_{u\in U} p_u \log\left(m_u^*(s)\right)$ in [23] to select utterances from multilingual sources. $\{p_u\}$ is the feature $u \in U$ distribution, and estimated from the development set [21, 22]. $l(s)$ is the length of utterance s. $m_u(s)$ measures the degree of feature u of the utterance s, and $m_u^*(S) = \sum_{s\in S}\frac{1}{l(s)} m_u(s)$ measures the average degree of feature u in the subset S normalized by the utterance length. The proposed submodular function considers the length normalization and also selects utterances which match the development set (the low-resource target language).

Although some phonemes could be shared among different languages, different languages have their own phoneme sets in general. Hence, phonetic related features which are language dependent cannot be used in submodular multilingual data selection. Gaussian mixture model (GMM) is widely used to capture the acoustic characteristics of utterances in speaker recognition and spoken term detection [32-34]. In addition, in our previous works [21-23], GMM-based tokenization has shown as good performance as the phonetic representation. In this paper, the Gaussian component index is used to represent the utterances from different languages. The Gaussian component index sequence of each utterance is converted to a vector space representation with term frequency-inverse document frequency (TF-IDF) weighting. Based on TF-IDF, $m_u(s)$ is computed according to equation $m_u(s) = tf(u,s)idf(u)$, where s is an utterance in a subset S, $tf(u,s)$ is the TF value for term u in utterance s, $idf(u)$ is the IDF value for term u. The term u is 2-gram of Gaussian component index in this paper. The greedy forward selection is then used to select utterances under the submodular optimization framework.

5. Experiments

5.1. Experimental setup

Four speech corpora (including Cantonese, Pashto, Turkish, and Tagalog) provided by the IARPA Babel program and the OpenKWS14 Tamil corpus were used in our KWS experiments. Each corpus contains both conversational speech and scripted speech. In the four Babel speech corpora, more than 100 hours of data are recorded in each full language pack (FLP), and each pronunciation lexicon only covers words appeared in the training transcription. When using the FLPs of the four languages to train a SHL-MDNN, both conversational and scripted speech were used. The very limited language pack (VLLP) of the Tamil corpus contains 3 hours of conversational telephone speech data, and its pronunciation lexicon only contains words that appear in the VLLP training transcription. The VLLP training transcription was used to train VLLP LMs, and the FLP training transcription was used to train FLP LMs in order to further verify our proposed approach. When using our proposed submodular data selection approach for updating the multilingual DNN, Tamil was viewed as the target language, and the other four languages were viewed as source languages.

All of keyword search experiments were carried out with the publicly available Kaldi toolkit [35], and the DNN training was performed on a single NVIDIA Tesla K20 GPU. The 22-dimensional fbank and 3-dimensional Kaldi pitch features were used to train the SHL-MDNN, and MFCC and Kaldi pitch features were extracted and used to train initial GMM-HMMs. The multilingual alignments were obtained using the language-dependent GMM-HMMs by force alignment to the utterances in the corresponding language, and the senone labels from the alignments were used to train the SHL-MDNN. Each source language dependent softmax layer contains about 4,500 senones. The target language softmax layer contains about 2,000 senones. The SHL-MDNN contains 6 hidden layers, and each hidden layer contains 1,500 units. α in Equation (1) was set to 0.1 so that the DNN update emphasized more on the target language.

The 15 hours of evaluation part 1 *Evalpart1* was used in our evaluation. The 10 hours of development set *Dev10h* was used to tune parameters. The evaluation keyword list provided by OpenKWS14 which contains 5,576 keywords or keyword phrases was used for evaluating keyword search systems. The actual term weighted value (ATWV) was used to measure the performance of keyword search [5].

5.2. Experimental results

5.2.1. Word based KWS experimental results

Table 1 shows the performance of different KWS systems when using the word-based VLLP LM and FLP LM for ASR decoding. Three hours of transcribed Tamil speech in FLP were used to train the acoustic models in these systems.

Table 1. *Performance of different KWS systems on Evalpart1. Word-based VLLP LM and FLP LM used in ASR decoding.*

System	Data Selection	Training / Update Time	ATWV	
			VLLP	FLP
Monolingual	N.A.	1 hours	0.1115	0.1494
Cross-lingual transfer	N.A.	3 hours	0.1250	0.1783
New SHL-MDNN training	N.A.	240 hours	0.1490	0.1998
SHL-MDNN update with new target language	Random	15 hours	0.1366	0.1840
SHL-MDNN update with new target language	Submodular	15 hours	0.1496	0.2027

Without leveraging the multilingual data, the monolingual baseline system performed the worst as expected. The ATWV of this system on *Dev10h* is 0.0765 and 0.1317 when the word-based VLLP LM and FLP LM are used respectively. On *Evalpart1*, the baseline system achieves the ATWV of 0.1115 and 0.1494 respectively.

We implemented two baseline approaches which leverage the multilingual data in acoustic model training. One is the cross-lingual model transfer approach [14], in which we re-estimated the top 3 layers (including the softmax layer) of the SHL-MDNN with the 3-hour transcribed Tamil speech. Its KWS performance is shown in the row labeled "Cross-lingual transfer" in Table 1. The other approach is to combine the transcribed Tamil data and all multilingual data to train a new SHL-MDNN from scratch. Its KWS performance is shown in

the row labeled "New SHL-MDNN training" in Table 1. Comparing with the cross-lingual model transfer approach, there are 19% and 12% relative ATWV improvements on *Evalpart1*. However, due to the large amount of multilingual data, it took a longer time to train the new SHL-MDNN model from scratch.

To examine our proposed data selection approaches for SHL-MDNN update, we implemented two approaches to select 10% all of multilingual data. One is our proposed submodular data selection, and the other one is random data selection. The selected multilingual data together with the target language data was used to update the original SHL-MDNN. The weighted cross-entropy criterion was also used to emphasize on the Tamil speech in the update. Table 1 shows the experimental results in the rows labeled "SHL-MDNN update with new target language". Comparing with the cross-lingual model transfer approach, our proposed approach achieves about 19% and 12% relative ATWV improvements on *Evalpart1* when the VLLP LM and FLP LM are used in ASR decoding respectively. Comparing with the system using the SHL-MDNN updated with the randomly selected multilingual data, our proposed approach achieves 10% relative ATWV improvements on Evalart1 when the VLLP LM and FLP LM are used in ASR decoding.

The model trained using our proposed approach has similar performance as the new SHL-MDNN model trained from scratch using the Tamil data and all multilingual data. However, our proposed approach is about 10 times faster than building the new SHL-MDNN from scratch as only 10% of multilingual data is used in our proposed approach. Another interesting finding from the experiments is that the multilingual data selected by our proposed approach does not contain Pashto data. The data proportion in our selected utterances for Cantonese, Turkish, and Tagalog is 21%, 32%, and 47% respectively. From the experiments, we can conclude that: (1) Not all multilingual data contributes to obtain a better model for the target language, and this is consistent with the observation in [9]; (2) Our proposed submodular data selection can select the multilingual data which is acoustically close to the target language data.

The weighted cross-entropy criterion is also important to improve the performance of KWS for our proposed approach. When using the equal weight in the SHL-MDNN update, ATWV is 0.1201 on *Evalpart1* when the VLLP LM is used, and it is similar to the system built by the cross-lingual model transfer approach. If the weight (α=0.1) is set to emphasize the target language, ATWV is 0.1496, which represents a 20% relative ATWV improvement.

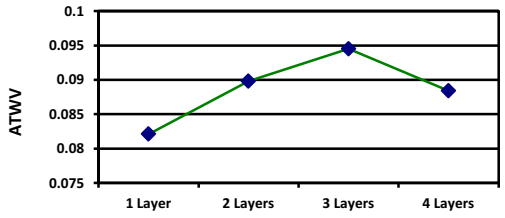

Fig. 1. *Performance of KWS on Dev10h by tuning different numbers of hidden layers for cross-lingual model transfer. Word based VLLP LM used in ASR decoding.*

In order to obtain better performance for the cross-lingual model transfer based KWS system, the cross-lingual DNN was tuned on *Dev10h*. Fig. 1 shows the KWS performance when different numbers of hidden layers are tuned in cross-lingual model transfer, and the word-based VLLP LM is used. In Fig. 1, "1 Layer" means that only the softmax layer is trained using the target language training data. "2 Layers" means that the first hidden layer of the SHLs (from top to bottom) and the softmax layer are re-estimated using the target language training data. In Fig. 1, we find that cross-lingual model transfer can help to improve the performance of KWS, and there are about 7.3%~23.5% relative ATWV improvements on *Dev10h*. The best ATWV on *Dev10h* was obtained when re-estimating "3 Layers" using the Tamil speech data.

5.2.2. Word-morph mixed KWS experimental results

Tamil is a morphologically rich language. The traditional word-based n-gram LM does not work well in this type of languages due to the huge number of different word forms. In the VLLP condition, the problem becomes even worse. Therefore, word-morph mixed language models and the corresponding word-morph mixed KWS systems were examined in our previous work [21]. Table 2 shows the performance of different KWS systems on *Evalpart1* when the word-morph mixed VLLP LM and FLP LM are used.

Table 2. *Performance of different KWS systems on Evalpart1. Word-morph mixed VLLP LM and FLP LM used in ASR decoding.*

System	Data Selection	ATWV	
		VLLP	FLP
Monolingual	N.A.	0.1115	0.1570
Cross-lingual transfer	N.A.	0.1304	0.1910
SHL-MDNN training	N.A.	0.1536	0.2009
SHL-MDNN update with new target language	Random	0.1379	0.1913
SHL-MDNN update with new target language	Submodular	0.1554	0.2038

By comparing Table 1 and Table 2, we can observe that the word-morph mixed LMs improve the performance of KWS. Using the word-morph mixed VLLP LM in ASR decoding, there are 0~4.3% relative ATWV improvements, and 0.5%~5.1% relative ATWV improvements are observed using the word-morph mixed FLP LM.

6. Conclusions

We propose a multilingual DNN update approach for low-resource keyword search. Using submodular data selection, a small amount of multilingual data is selected, and the selected multilingual data together the target language data are used to update the multilingual DNN acoustic model. Our proposed data selection approach can select multilingual data which is acoustically close to the target language. We illustrate that when conducting the multilingual DNN update, the weighted cross-entropy criterion is important to the final KWS performance. No matter which type of LM is used, our proposed approach outperforms the new SHL-DNN model trained using the target language and all multilingual data from scratch, which is the best baseline leveraging the multilingual data. Moreover, our proposed method greatly reduces the time to obtain the acoustic model which is as good as the best multilingual baseline. In the future, we will apply our proposed approach to other languages.

7. 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, pp.314-317.

[5] J. G. iscus, J. Ajot, J. S. Garofolo, G. Doddingtion, "Results of the 2006 Spoken Term Detection Evaluation," in *Proc. SIGIR 2007 Workshop on Searching Spontaneous Conversational Speech*, pp. 51–57.

[6] I. Szoeke, M. Fapso, and L. Burget, "Hybrid Word-subword Decoding for Spoken Term Detection," in *Proc. SIGIR* 2008, pp.121-129.

[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 Tamil," in *Proc. ICASSP* 2015, pp.5366-5370.

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

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

[11] 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 *ASRU* 2013, pp.138-143.

[12] 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.

[13] A. Ghoshal, P. Swietojanski, S. Renals, "Multilingual Training of Deep Neural Networks," in *Proc. ICASSP* 2013, pp.7319-7323.

[14] 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.

[15] P. Golik, Z. Tuske, R. Schluter, H. Ney, "Multilingual Features Based Keyword Search for Very Low-resource Languages," in *Proc. Interspeech* 2015, pp.1260-1264.

[16] Z. Tuske, P. Golik, D. Nolden, R. Schluter, H. Ney, "Data Augmentation, Feature Combination, and Multilingual Neural Networks to Improve ASR and KWS Performance for Low-resource Languages," in *Proc. Interspeech* 2014,pp.1420-1424.

[17] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, H. Bourlard, "Multilingual Deep Neural Network based Acoustic Modeling for Rapid Language Adaptation," in *Proc. ICASSP* 2014, pp.7639-7643.

[18] J. Cui, B. Kingsbury, B. Ramabhadran, A. Sethy, K. Audhkhasi, X. Cui, E. Kislal, L. Mangu, M. Nuubaum-Thom, M. Picheny, Z. Tuske, P. Golik, R. Schlüter, H. Ney, Mark J. F. Gales, K. M. Knill, A. Ragni, H. Wang, Philip C. Woodland, "Multilingual Representation for Low-resource Speech Recogntion and Keyword Search," *IEEE 2015 Workshop on Automatic Speech Recognition and Understanding*, pp.259-266.

[19] Q. B. Nguyen, J. Gehring, M. Muller, S. Stuker, A. Waibel, "Multilingual Shifting Deep Bottleneck Features for Low-resource ASR," in *Proc. ICASSP* 2014, pp.5607-5611.

[20] F. Grezl, M. Karafiat, and K. Vesely, "Adaptation of Multilingual Stacked Bottleneck Neural Network Structure for New Language," in Proc. ICASSP 2014, pp.7654-7658.

[21] 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, pp.4714-4718.

[22] 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, pp.4629-4633.

[23] C. Ni, C.-C. Leung, L. Wang, H. Liu, F. Rao, L. Lu, B. Ma, and H. Li, "Crosslingual Deep Neural Network based Submodular Unbiased Data Selection for Low-resource Keyword Search," in *Proc. ICASSP* 2016,pp.6015-6019.

[24] 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.

[25] 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.

[26] T. Schultz, and W. Waibel, "Language-independent and Language-adaptive Acoustic Model for Speech Recognition," Speech Communication, 2001, 35(1):31-51.

[27] H. Zheng, Z. Yang, L. Qiao, J. Li, and W. Liu, "Attribute Knowledge Integration for Speech Recognition Based on Multi-Task Learning Neural Networks," in *Proc. Interspeech* 2015, pp.543-547.

[28] R. Caruna, "Multitask Learning: A Knowledge-based Source of Inductive Bias," in *Proc. ICML* 1993, pp.41-48.

[29] M. L. Seltzer and J. Droppo, "Multi-task Learning in Deep Neural Networks for Improved Phoneme Recognition," *in Proc. ICASSP* 2013, pp. 6965-6969.

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

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

[32] T. Kinnunen, and H. Li, "An Overview of Text-independent Speaker Recognition: From Features to Supervectors," *Speech Communication*, 2010:52(1):12-40.

[33] Y. Zhang, "Unsupervised Speech Processing with Applications to Query-by-Example Spoken Term Detection," Ph.D Thesis.

[34] H. Wang, T. Lee, C.-C. Leung, B. Ma and H. Li, "Acoustic Segment Modeling with Spectral Clustering Methods," *IEEE Trans. On Audio, Speech and Language Processing*, 2015, 23(2):264-277.

[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," *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.