EFFICIENT METHODS TO TRAIN MULTILINGUAL BOTTLENECK FEATURE EXTRACTORS FOR LOW RESOURCE KEYWORD SEARCH

Chongjia Ni, Cheung-Chi Leung, Lei Wang, Nancy F. Chen, and Bin Ma

Institute for Infocomm Research (I^2R), A*STAR, Singapore

ABSTRACT

Training a bottleneck feature (BNF) extractor with multilingual data has been common in low resource keyword search. In a low resource application, the amount of transcribed target language data is limited while there are usually plenty of multilingual data. In this paper, we investigated two methods to train efficient multilingual BNF extractors for low resource keyword search. One method is to use the target language data to update an existing BNF extractor, and another method is to combine the target language data to train a new multilingual BNF extractor from the start. In these two methods, we proposed to use long short-term memory recurrent neural network based language identification to select utterances in the multilingual training data that are acoustically close to the target language. Experiments on Swahili in the OpenKWS15 data demonstrated the efficiency of our proposed methods. The first method facilitates rapid system development, while both methods outperform using baseline BNF extractors in terms of accuracy.

Index Terms**—**Keyword spotting, multilingual data selection, recurrent neural network, language identification

1. INTRODUCTION

In recent years, low resource keyword search (KWS) has gained much attention from researchers. Due to limitations of keyword-filler based KWS systems [1,2], such as high false alarm rates when large numbers of keywords or keyword phrases are involved, most state-of-the-art KWS systems are based on large vocabulary continuous speech recognition (LVCSR) [3-12].

Building state-of-the-art LVCSR systems requires large amounts of transcribed speech and linguistic knowledge of target languages. It is difficult and time-consuming to acquire these resources for some languages, especially for low resource languages. To overcome this limitation and improve the performance of KWS, researchers have proposed different methods for using the transcribed data from other languages to help to build acoustic models or feature extractors for low resource languages [13-23]. One method is to build multilingual DNN-HMM hybrid systems. Multilingual deep neural network (DNN) based cross-lingual knowledge transfer, as an effective method for transferring knowledge across languages, has attracted broad attention [13,15,17,18,21,22]. Another commonly used method is to extract bottleneck features from a multilingual DNN [14,16,19,20,23], and then the extracted features are used to train a GMM-HMM or DNN-HMM recognizer. Some studies have shown that multilingual deep bottleneck features are more efficient than multilingual DNN based cross-lingual knowledge transfer [24,25]. There have also been some studies shown that not all multilingual data can contribute equally to the KWS performance of target languages when using multilingual data to train bottleneck feature extractors or DNN-HMM hybrid systems for cross-lingual knowledge transfer [13, 26]. Thus, it is better to select subsets of multilingual data for building efficient multilingual DNN models or bottleneck feature extractors for target languages.

In this paper, we investigate two methods to build efficient multilingual BNF extractors for KWS of a low resource target language. In these two methods, we use long short-term memory (LSTM) recurrent neural network (RNN) based language identification (LID) to select utterances in the multilingual training data that are acoustically close to the target language. The first method is to use these two sets of data to update an existing BNF extractor, and the second method is to combine the target language data and the selected multilingual data to train a new multilingual BNF extractor from the start. These two methods are aimed at reducing the amount of multilingual data used for training, and thus reducing the time taken to build new multilingual BNF extractors for the target language. To our knowledge, these two methods have never been compared.

Note that Zhang *et al.* and Chuangsuwanich *et al.* have proposed to use LID for multilingual data selection [13, 25]. However, our present work differs from this work in the following aspects: (1) While [13,25] only consider to train a multilingual BNF extractor from the start, our work also considers updating an existing multilingual BNF extractor using the selected data. We will demonstrate that rapidly updating an existing BNF extractor can have comparable performance to a new BNF extractor built from scratch. (2) [13,25] do not consider including target language data when training a multilingual BNF extractor. In addition to selecting multilingual data based on LID, it is worth noting that our previous work [26] considers updating a shared-hidden-layers multilingual DNN hybrid system with the multilingual data section using a submodular method with term frequency-inverse document frequency of frame based Gaussian component index [8,11,12].

The rest of the paper is organized as follows. A multilingual data selection method by using long short-term memory recurrent neural network based language identification is presented in Section 2. In Section 3, methods for building multilingual deep bottleneck extractors are introduced. The experimental setup, results and analysis are presented in details in Section 4. Section 5 concludes the paper.

2. MUTLTILINGUAL DATA SELECTION USING LONG SHORT-TERM MEMORY RECURRENT NEURAL NETWORK BASED LANGUAGE IDENTIFICATION

Long short-term memory (LSTM) recurrent neural network (RNN) has been applied to language identification (LID), and has obtained state-of-the-art performance probably due to its more sophisticated modeling of long-term information in speech [27]. In this paper, we propose to use LSTM RNN to select the utterances in the multilingual training data that are misclassified into the target language with high probability, to train BNF extractors. The reason for this is that if an utterance in a language has high probability to be misclassified into another language, it is reasonable to say that

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

these two languages are similar. Being similar is possibly due to the effect from recording environments or/and speaker characteristics.

When building an LSTM RNN model for LID, we define $\mathrm{N}+1$ unique frame labels, where N is the number of languages (including the target language). For each language, the speech frames of the utterances are labeled using the language identity index. An extra label is used to represent silence frames of utterances in all languages. Let S be the set of utterances, $\mathrm{x} = (\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_\mathrm{T})$ be an input utterance feature, $\mathrm{p}(\mathrm{L}|\mathrm{x}_\mathrm{t}) = \left(\mathrm{p}(\mathrm{l}_0|\mathrm{x}_\mathrm{t}), \mathrm{p}(\mathrm{l}_1|\mathrm{x}_\mathrm{t}), \cdots, \mathrm{p}(\mathrm{l}_\mathrm{N}|\mathrm{x}_\mathrm{t})\right)$ be the posterior vector for input feature x_t at frame t, where $\mathrm{p}(\mathrm{l}_\mathrm{i}|\mathrm{x}_\mathrm{t})$ is the posterior of language l_i for input feature x_t, l_0 is the silence output target index, a function can be defined as follows:

$$f(S) = \sum_{s\in S} \log\left(Proj_{i_k}\left(\frac{1}{T(s)}\sum_{t=1}^{T} p(L|x_t^s)\right)\right) \qquad (1)$$

where $T(s)$ is the number of frames of utterance s, x_t^s is the t^{th} input feature of the input utterance s, $Proj_{i_k}(\cdot)$ is the projection function in order to get the i_k component for a vector, and i_k is the target language index. In order to reduce frame level misclassification, the posterior probability is averaged for each utterance. When using Eq. (1) for multilingual data selection, the utterances with high misclassification probability values are selected.

3. METHODS TO BUILD MULTILINGUAL DEEP BOTTLENECK EXTRACTORS

A shared-hidden-layer multilingual deep neural network (SHL-MDNN) [18] is used as a feature extractor in this paper. The extracted BNFs are then used to train a GMM-HMM or DNN-HMM recognizer. In this multilingual network, hidden layers are shared across many languages while the softmax layers are language dependent. An internal bottleneck layer is used to extract multilingual BNFs, which carry information for phonetic discrimination in multiple source languages. This BNF extractor facilitates the use of resources from other languages. Multilingual training data are selected according to the average misclassification posterior probability at utterance level according to Eq. (1).

Note that we need to consider the number of source languages involved in these selected utterances. More selected source languages increase the size of the output senone layer, which increases the time taken to train the multilingual BNF extractor. For convenience, we filter out some source languages in which not many utterances are selected. After that, we can update an existing BNF extractor, or train a new multilingual BNF extractor from the start, together with the target language data.

3.1 Rapid update of existing multilingual BNF extractors

The performance of KWS can be improved by updating an existing multilingual hybrid DNN for a target language. Due to the limited amount of the target language data, only the new softmax layer (output nodes in the new softmax layer corresponding to the senones created for the target language) is trained while the parameters in the shared hidden layers are fixed. Our work in [26] shows that the system improvement is not satisfactory when this multilingual network is used for acoustic modeling. To further improve the system, a small amount of multilingual data, which is acoustically close to the target language, is selected and used together with the target language data to update the parameters in the existing multilingual network.

This motivates us to update an existing multilingual BNF extractor for a new target language. In this paper, we also consider the situation that more data from different languages are available when a new target language is identified, and a multilingual BNF extractor is readily available before the target language is identified. When updating an existing MDNN, a weighted cross-entropy criterion is used in [26]. A higher weight is used to emphasize the target language. In this paper, we also tried a higher weight to emphasize the target language in the weighted cross-entropy criterion in the training of multilingual BNF extractors, but no significant gain was observed, so equal weighting is used in this work. We believe that the weighting used in training a BNF extractor is not as sensitive as that used in training a hybrid DNN-HMM for acoustic modeling [26].

3.2 Training from scratch with selected multilingual data

The first method is not optimized for the target language because: (1) the training data used for training an existing BNF extractor may not contribute (or even hurt) the performance for the target language; and (2) it is difficult to efficiently adapt an existing BNF (trained by lots of multilingual data) with a small amount of target language data. An intuitive idea is to train a new multilingual BNF extractor from the start together with the target language data.

Previous research has demonstrated that not all multilingual data can contribute equally to the final KWS performance for the target language [13, 26], and the utterances that are acoustically similar to the target language, are more useful for building a multilingual BNF extractor for the target language. Therefore, similar to [13, 25], in this paper, when building the BNF extractor from the start, we proposed to use the LSTM RNN LID-based multilingual data selection method to select utterances from different languages for building an efficient multilingual BNF extractor. Since the target language data should be the most efficient data to train the BNF extractor for the target language, we consider including the target language data in the training procedure. Equal weight is used in cross-entropy criterion to build the multilingual BNF extractor.

4. EXPERIMENTS

4.1. Experimental setup

The OpenKWS15 Swahili data provided by the IARPA Babel program as the target language was used in the experiments. The other languages, a total of 23 languages, released by the IARPA Babel program in three stages were also used in our experiments for language identification model training and cross-lingual transfer.

In the data of each language, there are three different conditions, including full language pack (FLP), limited language pack (LLP), and very limited language pack (VLLP), which correspond to different training data sets:

FLP: 60-170 hours of transcribed speech, depending on which phase of the Babel program the data was designed for, and the corresponding FLP pronunciation lexicon.

LLP: a subset of 10 hours of transcribed speech in FLP, and the corresponding LLP pronunciation lexicon.

VLLP: a subset of 3 hours of transcribed speech in FLP.

A number of data sets used for training monolingual or multilingual BNF extractors in our experiments were described as follows:

VLLP-TL: the 3 hours of target language data in VLLP.

Baseline-Multilingual-509h: 509 hours (Cantonese: 175.2 hours, Pashto: 111.1 hours, Turkish: 107.4 hours, Tagalog: 115.7 hours) of data randomly selected from FLP of 4 languages. We assume that a multilingual BNF extractor built using this set of data is available before the target language is identified.

Baseline-Multilingual-14h-LID: 14 hours (3.5 hours from each language) of data selected from "**Baseline-Multilingual-509h"**, have the highest misclassification probabilities defined according to (1) in each language.

Baseline-Multilingual-14h-Sub: 14 hours (3.5 hours from each language) of data selected from "**Baseline-Multilingual-509h**" using our previously proposed submodular method in [26].

Proposed-Multilingual-96h: 96 hours of data (Haitian Creole: 29.7 hours, Zulu: 21.6 hours, Dholuo: 23.9 hours, Vietnamese: 20.7 hours) selected by our proposed LID based data selection method, which have the highest misclassification probabilities defined according to (1).

Proposed-Multilingual-14h: 14 hours (3.5 hours from each language) of data selected from "**Proposed-Multilingual-96h**", which have the highest misclassification probabilities defined according to (1) in each language.

Creole-14h: 14 hours of Haitian Creole data selected from "**Proposed-Multilingual-96h**", which have the highest misclassification probabilities defined according to (1).

Submodular-Multilingual-96h: 96 hours of data (Zulu: 20.1 hours, Pashto: 35.0 hours, Vietnamese: 27.6 hours, Cantonese: 13.3 hours) selected using our previously proposed submodular method in [26].

In order to fairly compare with the baseline multilingual data set, we constrain only four languages of data to be selected when using LID or submodular method to select utterances.

The web text data of the target language provided by the IARPA Babel program was also used to train a 3-gram language model. The web text based LM was interpolated with the 3-gram LM trained by using the VLLP training transcription. The interpolation weight was optimized by minimizing the perplexity on the transcription of 10 hours of development set *Dev10h*. The interpolated LM is denoted as "Web-data LM". The VLLP training transcription based LM is denoted as "Training transcription LM".

To evaluate our proposed methods, we built different KWS systems. When building these KWS systems, the Kaldi toolkit and corresponding KWS recipe with the same settings were used in order to fairly compare the performance of different KWS systems [28]. The features used for training a monolingual BNF extractor or multilingual BNF extractor are 22 fbank features and 3 Kaldi pitch related features and their delta and delta-delta features (fbank+pitch+Δ+ΔΔ). After extracting 42 BNF features, they were concatenated with the 75 fbank+pitch+Δ+ΔΔ features, and then 117 fbank+pitch+Δ+ΔΔ+BNF features were obtained. These features were used in the training of a target language hybrid DNN (6 hidden layers, 1,024 hidden units for each hidden layer). It is first trained based on cross-entropy criterion, and then based on sMBR criterion for sequence training. The alignments used for training hybrid DNN model were generated by a discriminative trained GMM-HMM system. The number of senones of the target language is 2,027. Only the cross-entropy criterion was used for multilingual deep bottleneck extractor (6 hidden layers, except the bottleneck layer includes 42 hidden units, other hidden layers include 1500 hidden units) training.

The keyword list provided for OpenKWS15, which contains 4,454 keywords or keyword phrases, was used to evaluate different keyword search systems. All KWS systems are word based KWS systems. When "Web-data LM" is used for ASR decoding, the number of out-of-vocabulary (OOV) keywords or keyword phrases is 260. When using "Training transcription LM" for ASR decoding, the number of OOV keywords or keyword phrases is 2,667. The actual term weighted value (ATWV) and word error rate (WER) were used to measure the performance of KWS systems and the underlying ASR systems. The 15 hours of evaluation part 1 *Evalpart1* was used for our evaluation.

4.2 Experimental results

In general, the web-data LM improved the performance of all KWS systems and their underlying ASR systems and we observed that if a system performs well with the web-data LM, it also performs well with training transcription LM. Table 1 lists the performance of different baseline keyword search systems. From Table 1, we can see that the baseline multilingual BNFs provide a 17.5% relative ATWV improvement compared with the monolingual BNFs when "Web-data LM" is used. Though the baseline multilingual BNFs improved the KWS performance, it could be improved further if the multilingual BNF extractor was trained using more data that were acoustically close to the target language.

Table 1. Performance of baseline KWS systems on *Evalpart1.*

BNF extractor	Data set for training BNF extractor	Web-data LM		Training transcription LM	
		WER	ATWV	WER	ATWV
Baseline Monolingual	VLLP-TL	67.4	0.308	69.3	0.194
Baseline Multilingual	Baseline-Multilingual-509h	64.5	0.361	69.0	0.216

Table 2 shows the performance on *Evalpart1* when different data sets were used to rapidly update the baseline multilingual BNF extractor.

Table 2. The performance of different KWS systems on *Evalpart1* by rapidly updating the baseline multilingual BNF extractor using 17 hours of multilingual data.

BNF extractor	Data set for updating BNF extractor	Web-data LM		Training transcription LM	
		WER	ATWV	WER	ATWV
R1	Baseline-Multilingual-14h-LID + VLLP-TL	62.1	0.396	66.7	0.239
R2	Baseline-Multilingual-14h-Sub + VLLP-TL	62.3	0.390	67.1	0.238
R3	Proposed-Multilingual-14h + VLLP-TL	**61.4**	**0.397**	**66.0**	**0.242**
R4	Creole-14h + VLLP-TL	61.6	0.389	66.3	0.231

As the target language data is the most efficient data for building a multilingual BNF extractor targeted for the target language, the target language data was always included to rapidly update the baseline multilingual BNF extractor. We observed that: (1) Updating an existing BNF extractor provided significant improvement. Comparing with the baseline multilingual BNFs, relative ATWV improvements between 7.8% and 10.0% were obtained when "Web-data LM" was used. Comparing the baseline multilingual BNF extractor and extractor R1 further confirms that not all multilingual data contribute to the performance and the performance improvement could be achieved while less than 3% of data in "Baseline-Multilingual-509h" was used in the update. This kind of method is particularly suitable to the situation when rapid system development is needed. (2) The kind of method worked well even if a new set of multilingual data was involved to update a readily available multilingual BNF extractor (as shown in R3). (3) The utterances selected by our proposed LID based method showed slightly better performance by comparing R1 and R2. (4) Selecting utterances from more languages (R3) provided slightly better performance than selecting the same amount of utterances from one language (R4).

Table 3. The performance of different KWS systems on *Evalpart1* by training multilingual BNF extractors from the start.

BNF extractor	Data set for training BNF extractor	Web-data LM		Training transcription LM	
		WER	ATWV	WER	ATWV
S1	Baseline-Multilingual-509h + VLLP-TL	61.2	0.413	65.7	0.243
S2	Proposed-Multilingual-96h	60.9	0.407	65.6	0.239
S3	Proposed-Multilingual-96h + VLLP-TL	**60.7**	**0.416**	**65.6**	**0.244**
S4	Submodular-Multilingual-96h	61.3	0.399	65.8	0.237
S5	Submodular-Multilingual-96h + VLLP-TL	61.1	0.402	65.7	0.237
S6	Creole-14h + VLLP-TL	65.1	0.372	69.5	0.221

Table 3 shows the performance on *Evalpart1* when different data sets were used to train a new multilingual BNF extractor from scratch. We observed that: (1) Training a new BNF extractor with our proposed data selection method (S2 and S3) provided significant improvement. Comparing with the baseline multilingual BNFs, relative ATWV improvements between 3.0% and 15.2% were obtained when "Web-data LM" was used. Moreover, although less than 20% of data in "Baseline-Multilingual-509h" was selected to train a new BNF extractor, a slightly better improvement was obtained using the extractor S3. (2) Although adding the target language data to "Baseline-Multilingual-509h" data to train a new extractor provided a 14% relative ATWV improvement, the relative improvement dropped to 2% when the target language data are added to "Proposed-Multilingual-96h" data (comparing S2 and S3). The improvement was similarly diminished when the target language data were added to "Proposed-Multilingual-96h" data. (3) The utterances selected using our proposed LID based method showed slightly better performance than those selected by the previously proposed submodular method (comparing S2 and S4, and S3 and S5). (4) The amount of selected data also affects the performance of the BNF extractor. When only the 14 hours of Creole data, which are acoustically closest to the target language data, were added to the target language data to train a new extractor (S6), the extractor was obviously not as good as the extractor trained with 96 hours of acoustically close data (S3) though the extractor S6 still outperformed the baseline monolingual BNF extractor.

4.3 Experimental analysis

As mentioned in Section 4.2, only a slightly better improvement was obtained when adding the target language data to "Proposed-Multilingual-96h" (comparing S2 and S3), which is in contrast to the obvious gain when adding the data to "Baseline-Multilingual-509h" (comparing S1 and baseline multilingual BNF extractor). Fig. 1 shows the similarity measure between different source languages and the target language data based on the LSTM RNN trained for LID. It is a posterior probability measurement averaged over all the utterance frames of each source language. A high value means that the misclassified language is more similar to the target language.

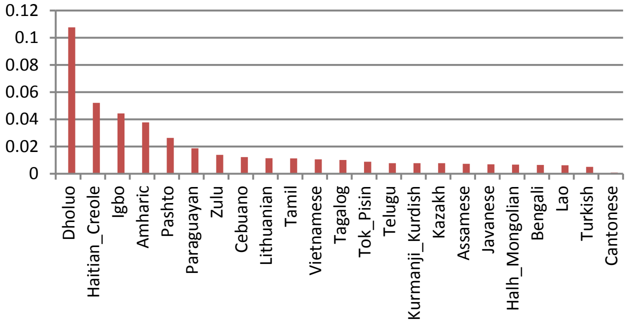

Fig. 1. Similarity measure between different source languages and target language (Swahili). The vertical axis denotes the average misclassification posterior probability of all utterance frames of each language.

From Fig. 1, we can find that the languages in the "Proposed-Multilingual-96h" data are more similar to the target language than the languages in "Baseline-Multilingual-509h" data. Since more data acoustically close to the target language data were selected in "Proposed-Multilingual-96h", it is reasonable that the extra gain was not obvious when the target language data was further added to the training of the proposed multilingual bottleneck extractor.

It is worth noting that the "Proposed-Multilingual-96h" data were formed by the first 96 hours of utterances with the highest utterance-level (not language-level) misclassification probability. We found that only two languages (Haitian Creole and Dhohuo) in "Proposed-Multilingual" data were ranked the most similar languages to Swahili. We believe that some utterances in Zulu and Vietnamese not selected in "Proposed-Multilingual" are not acoustically close (possibly due to the effect from recording environments or/and speakers) to the Swahili data, so that the similarity measures between these two languages and the Swahili data as shown in Fig. 1 are not among the highest four.

4. CONCLUSIONS

In this paper, we studied how to efficiently train BNF extractors. In order to select multilingual data to efficiently build deep bottleneck extractors, we proposed a novel multilingual data selection method. The proposed method can select utterances that are acoustically similar to the target language data. Experimental results showed that the selected multilingual data were more helpful for building efficient multilingual deep bottleneck extractors though a small portion of multilingual data was used. In this paper, rapidly updating an existing bottleneck extractor and training a new multilingual bottleneck extractor from scratch were investigated. We observed that when combining target language data and relatively un-similar multilingual data to build a BNF extractor, the gain was obvious for KWS of the target language. Meanwhile, when combining target language data and relatively similar multilingual data to build a BNF extractor, the gain was not as obvious as the previous case.

5. 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*, vol. 38, no. 11 pp. 1870-1878, 1990.

[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. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddintion, "Results of the 2006 Spoken Term Detection Evaluation," in *Proc. ACM SIGIR* 2007, Workshop in Searching Spontaneous Conversational Speech (SSCS 2007), pp. 51-56.

[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, 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, pp. 5366-5370.

[8] 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," in *Proc. ICASSP* 2016, pp. 6015-6019.

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

[10] N. F. Chen, H. Xu, X. Xiao, V. H. Do, C. Ni, I.-F. Chen, S. Sivadas, C.-H. Lee, E. S. Chng, B. Ma, H. Li, "Exemplar-inspired Strategies for Low-resource Spoken Keyword Search in Swahili," in *Proc. ICASSP* 2016, pp. 6040-6044.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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," in *Proc. ASRU* 2015, pp. 259-266.

[23] 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.

[24] H. Xu, V. H. Do, X. Xiao, and E. S. Chng, "A Comparative Study of BNF and DNN Multilingual Training on Cross-lingual Low-resource Speech Recognition," in *Proc. Interspeech* 2015, pp. 2132-2136.

[25] E. Chuangsuwanich, Y. Zhang, J. Glass, "Multilingual Data Selection for Training Stacked Bottleneck Features," in *Proc. ICASSP* 2016, pp. 5410-5414.

[26] C. Ni, L. Wang, C.-C. Leung, F. Rao, L. Lu, B. Ma, and H. Li, "Rapid Update of Multilingual Deep Neural Network for Low-resource Keyword Search," in *Proc. Interspeech* 2016, pp. 3698-3702.

[27] J. Gonzalez-Dominguez, I. Lopez-Moreno, H. Sak, J. Gonzalez-Rodriguez, P. J. Moreno, "Automatic Language Identification using Long Short-Term Memory Recurrent Neural Networks," in *Proc. Interspeech* 2014, pp. 2155-2159.

[28] 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 *Proc. ASRU* 2011.